А.В.Блохин

ТЕОРИЯ ЭКСПЕРИМЕНТА

Курс лекций в двух частях

Часть 2

Блохин А.В. Теория эксперимента [Электронный ресур]: Курс лекций в двух частях: Часть 2. — Электрон. текст. дан. (1,0 Мб). — Мн.: Научно-методический центр "Электронная книга БГУ", 2003. — Режим доступа: http://anubis.bsu.by/publications/elresources/Chemistry/blohin2.pdf . — Электрон. версия печ. публикации, 2002. — PDF формат, версия 1.4 . — Систем. требования: Adobe Acrobat 5.0 и выше.

МИНСК

«Электронная книга БГУ»

2003

© Блохин А.В.
© Научно-методический центр «Электронная книга БГУ»
www.elbook.bsu.by
elbook@bsu.by

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ХИМИЧЕСКИЙ ФАКУЛЬТЕТ
Кафедра физической химии

А. В. Блохин

ТЕОРИЯ ЭКСПЕРИМЕНТА

Курс лекций

В двух частях

Часть 2

МИНСК
2002

УДК 542(042)
ББК 24.в.я73
Б70

Рецензенты:

кандидат химических наук, доцент *Н. И. Горошко;*
старший преподаватель
кафедры физической химии БГУ *Л. М. Володкович*

Печатается по решению
Редакционно-издателъского совета
Белорусского государственного университета

Блохин А. В.
Б70 Теория эксперимента: Курс лекций. В *2* ч. Ч. 2 / А, В. Блохин. - Мн.: БГУ, 2002. - 67 с.
ISBN 985-445-815-6

Вторая часть курса лекций посвящена статистическим методам оптимизации экспериментальных исследований в физической химии и содержиг основы методов регрессионного, корреляционного и дисперсионною анализов и планирования экстремальною эксперимента.
Предназначено для студентов IV курса химического факультета.

УДК 542(042)
ББК 24.в.я73

ISBN 985-445-815-6(ч. 2)
ISBN 985-445-792-3

@Блохин А. В., 2002
@БГУ, 2002

СОДЕРЖАНИЕ

ЛЕКЦИЯ 7 ____ *4*

7.1. Системы случайных величин. Функция и плотность распределения системы двух случайных величин. Условные законы распределения ____ 4

7.2. Стохастическая связь. Ковариация. Коэффициент корреляции. Регрессия ____ 6

7.3. Выборочный коэффициент корреляции. Проверка гипотезы об отсутствии корреляции ____ 9

7.4. Приближенная регрессия. Метод наименьших квадратов ____ 13

ЛЕКЦИЯ 8 ____ *16*

8.1. Линейная регрессия от одного параметра ____ 16

8.2. Регрессионный анализ ____ 17

8.2.1. Проверка адекватности приближенного уравнения регрессии эксперименту ____ 18

8.2.2. Оценка значимости коэффициентов уравнения регрессии ____ 19

8.2.3. Оценка доверительного интервала для искомой функции ____ 21

8.3. Оценка тесноты нелинейной связи ____ 22

8.4. Аппроксимация. Параболическая регрессия ____ 23

8.5. Приведение некоторых функциональных зависимостей к линейному виду ____ 24

8.6. Метод множественной корреляции ____ 26

ЛЕКЦИЯ 9 ____ *28*

9.1. Задачи дисперсионного анализа. Однофакторный дисперсионный анализ ____ 28

9.2. Двухфакторный дисперсионный анализ ____ 32

ЛЕКЦИЯ 10 ____ *37*

10.1. Планирование эксперимента при дисперсионном анализе ____ 37

10.2. Постановка задачи при планировании экстремальных экспериментов ____ 40

10.3. Полный факторный эксперимент типа 2^2: матрица планирования, вычисление коэффициентов уравнения регрессии ____ 42

ЛЕКЦИЯ 11 ____ *45*

11.1. Матрица планирования полного факторного эксперимента типа 23 ____ 45

11.2. Проверка значимости коэффициентов и адекватности уравнения регрессии, полученных при обработке результатов ПФЭ 2^2 и 2^3 ____ 47

11.3. Дробный факторный эксперимент. Планы типа $2^{k\text{-}1}$ ____ 50

ЛЕКЦИЯ 12 ____ *54*

12.1. Оптимизация методом крутого восхождения по поверхности отклика ____ 54

12.2. Описание функции отклика в области, близкой к экстремуму. Композиционные планы Бокса-Уилсона ____ 56

12.3. Ортогональные планы второго порядка, расчет коэффицентов уравнения регрессии ____ 59

12.4. Метод последовательного симплекс-планирования ____ 62

ЛЕКЦИЯ 7

Системы случайных величин. Функция и плотность распределения системы двух случайных величин. Условные законы распределения. Стохастическая связь. Ковариация. Коэффициент корреляции, его свойства. Линии регрессии. Выборочный коэффициент корреляции; проверка гипотезы об отсутствии корреляции. Приближенная регрессия; метод наименьших квадратов.

7.1. Системы случайных величин. Функция и плотность распределения системы двух случайных величин. Условные законы распределения

На практике чаще всего приходится иметь дело с экспериментами, результатом которых является не одна случайная величина, а две и более, образующие систему. Свойства системы случайных величин не ограничиваются свойствами величин, в нее входящих; они определяются также взаимосвязью (зависимостями) этих случайных величин. Информация о каждой случайной величине, входящей в систему, содержится в ее законе распределения.

Рассмотрим систему из двух случайных величин X и Y. Функцией распределения такой системы называется вероятность совместного выполнения двух неравенств

$$F(x,y) = P\left(X < x, Y < y\right). \qquad (7.1)$$

Плотность распределения системы $f(x, y)$ определяется как вторая смешанная производная $F(x, y)$

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x\, \partial y}. \qquad (7.2)$$

Вероятность попадания точки (X, Y) в произвольную область D равна

$$P[(X,Y) \subset D] = \iint\limits_{(D)} f(x,y)\,\mathrm{d}x\,\mathrm{d}y. \qquad (7.3)$$

Свойства плотности распределения:

1) она является неубывающей функцией:

$$f(x,y) \geq 0; \qquad (7.4)$$

2) вероятность попадания случайной точки на всю координатную плоскость равна вероятности достоверного события:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,dx\,dy = 1; \tag{7.5}$$

3) функция распределения выражается через плотность распределения как

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y)\,dx\,dy; \tag{7.6}$$

4) плотность распределения каждой из случайных величин можно получить следующим образом:

$$F_1(x) = F(x,\infty) = \int_{-\infty}^{x}\int_{-\infty}^{+\infty} f(x,y)\,dx\,dy, \tag{7.7}$$

$$f_1(x) = \frac{dF_1(x)}{dx} = \int_{-\infty}^{+\infty} f(x,y)\,dy, \tag{7.8}$$

$$f_2(y) = \frac{dF_2(y)}{dy} = \int_{-\infty}^{+\infty} f(x,y)\,dx. \tag{7.9}$$

Чтобы полностью охарактеризовать систему (т. е. получить ее закон распределения), кроме распределения каждой величины, входящей в систему, необходимо знать и связь между этими величинами. Эта зависимость характеризуется с помощью *условных законов распределения*.

Условным законом распределения величины Y, входящей в систему (X, Y), называется ее закон распределения при условии, что другая случайная величина X приняла определенное значение x. Условная функция распределения обозначается $F(y/x)$, плотность распределения — $f(y/x)$. Для условных плотностей распределений справедлива *теорема умножения законов распределения*:

$$f(x,y) = f_1(x)\,f(y/x), \tag{7.10}$$

$$f(x,y) = f_2(y)\,f(x/y). \tag{7.11}$$

Тогда

$$f(y/x)=\frac{f(x,y)}{f_1(x)}=\frac{f(x,y)}{\int\limits_{-\infty}^{\infty} f(x,y)\,dy}, \qquad (7.12)$$

$$f(x/y)=\frac{f(x,y)}{f_2(y)}=\frac{f(x,y)}{\int\limits_{-\infty}^{\infty} f(x,y)\,dx}. \qquad (7.13)$$

7.2. Стохастическая связь. Ковариация. Коэффициент корреляции. Регрессия

Стохастической связью между случайными величинами называется такая связь, при которой с изменением одной величины меняется распределение другой. *Функциональной зависимостью* называется такая связь между случайными величинами, при которой при известном значении одной из величин можно точно указать значение другой.

В отличие от функциональной связи при стохастической связи с изменением величины X величина Y имеет лишь тенденцию изменяться. По мере увеличения тесноты стохастической зависимости она все более приближается к функциональной, а в пределе ей соответствует. Крайняя противоположность функциональной связи — полная независимость случайных величин.

Если случайные величины независимы, то согласно теореме умножения (7.10–7.11) получаем

$$f(y/x)=f_2(y) \text{ и } f(x/y)=f_1(x), \qquad (7.14)$$

$$f(x,y)=f_1(x)f_2(y). \qquad (7.15)$$

Условие (7.15) можно использовать в качестве необходимого и достаточного критерия независимости двух случайных величин, если известны плотности распределения системы и случайных величин, в нее входящих.

При неизвестном законе распределения системы для оценки тесноты стохастической связи чаще всего используется *коэффициент корреляции*. Дисперсия суммы двух случайных величин X и Y равна

$$D\{X+Y\}=M\left\{[X+Y-M(X+Y)]^2\right\}=M\left\{[X-M(X)+Y-M(Y)]^2\right\}=$$

$$=M[X-M(X)]^2+2M\{[X-M(X)][Y-M(Y)]\}+M[Y-M(Y)]^2=$$

$$= D(X) + 2M\{[X - M(X)][Y - M(Y)]\} + D(Y). \qquad (7.16)$$

Если X и Y независимы, то

$$D(X+Y) = D(X) + D(Y).$$

Тогда зависимость между X и Y существует, если

$$M\left([X - m_x][Y - m_y]\right) \neq 0. \qquad (7.17)$$

Величина (7.17) называется *корреляционным моментом*, или *ковариацией* $\mathrm{cov}\{XY\}$, (cov_{xy}) случайных величин. Она характеризует не только зависимость величин, но и их рассеяние.

Из (7.17) следует, что если одна из величин мало отклоняется от своего математического ожидания, то ковариация будет мала даже при тесной стохастической связи. Чтобы избежать этого, для характеристики связи используют безразмерную величину, называемую *коэффициентом корреляции*:

$$r_{xy} = \frac{\mathrm{cov}_{xy}}{\sigma_x \sigma_y} = \frac{M\left([X - m_x][Y - m_y]\right)}{\sigma_x \sigma_y}, \qquad (7.18)$$

где σ_x и σ_y — стандартные отклонения X и Y.

Случайные величины, для которых ковариация (значит, и коэффициент корреляции) равна нулю, называются *некоррелированными*. Равенство нулю коэффициента корреляции не всегда означает, что случайные величины X и Y независимы: связь может проявляться в моментах более высокого порядка (по сравнению с математическим ожиданием). Только в случае нормального распределения при $r_{xy} = 0$ связь между случайными величинами однозначно отсутствует.

Плотность нормального распределения системы двух случайных величин выражается следующей формулой:

$$f(x,y) = \frac{1}{2\pi \sigma_x \sigma_y \sqrt{1 - r^2}} \times,$$

$$\times \exp\left\{ -\frac{1}{2(1 - r^2)} \left[\frac{(x - m_x)^2}{\sigma_x^2} - \frac{2r(x - m_x)(y - m_y)}{\sigma_x \sigma_y} + \frac{(y - m_y)^2}{\sigma_y^2} \right] \right\}, \qquad (7.19)$$

где r — коэффициент корреляции. Если X и Y некоррелированы (т. е. $r = 0$), то из (7.19) следует, что

$$f(x,y) = \frac{1}{2\pi \sigma_x \sigma_y} \exp\left\{ -\frac{1}{2} \left[\frac{(x - m_x)^2}{\sigma_x^2} + \frac{(y - m_y)^2}{\sigma_y^2} \right] \right\} =$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_x} \exp\left[-\frac{(x-m_x)^2}{2\sigma_x^2}\right] \frac{1}{\sqrt{2\pi}\,\sigma_y} \exp\left[-\frac{(y-m_y)^2}{2\sigma_y^2}\right] =$$

$$= f_1(x) f_2(y), \qquad (7.20)$$

т. е. нормально распределенные случайные величины X и Y не только некоррелированы, но и независимы.

Отметим следующие свойства коэффициента корреляции:

1) величина r_{xy} не меняется от прибавления к X и Y неслучайных слагаемых;
2) величина r_{xy} не меняется от умножения X и Y на положительные числа;
3) если одну из величин, не меняя другой, умножить на –1, то на –1 умножится и коэффициент корреляции.

Тогда, если от исходных величин перейти к нормированным

$$X_0 = \frac{X - m_x}{\sigma_x},\ Y_0 = \frac{Y - m_y}{\sigma_y},$$

величина r_{xy} не изменится: $r_{x_0 y_0} = r_{xy}$. Из (7.16) и (7.18) следует, что

$$\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y) + 2r_{xy}\sqrt{\sigma^2(X)\sigma^2(Y)}. \qquad (7.21)$$

Для нормированных величин $\sigma^2(X_0) = \sigma^2(Y_0) = 1$, тогда

$$\sigma^2(X_0 + Y_0) = 2 + 2r_{xy}. \qquad (7.22)$$

Аналогично в случае разности $(X - Y)$ можно получить, что

$$\sigma^2(X_0 - Y_0) = 2 - 2r_{xy}. \qquad (7.23)$$

По определению дисперсии

$$\sigma^2(X_0 + Y_0) \ge 0 \text{ и } \sigma^2(X_0 - Y_0) \ge 0,$$

следовательно

$$2 + 2r_{xy} \ge 0,\ 2 - 2r_{xy} \ge 0,$$

$$r_{xy} \ge -1,\ r_{xy} \le 1,$$

$$-1 \le r_{xy} \le 1. \qquad (7.24)$$

При $r_{xy} = \pm 1$ имеем линейные функциональные зависимости вида

$$y = b_0 + b_1 x,$$

при этом если $r_{xy} = 1$, то $b_1 > 0$; если $r_{xy} = -1$, то $b_1 < 0$.

Если между величинами X и Y имеется произвольная стохастическая связь, то $-1 < r_{xy} < 1$. При $r_{xy} > 0$ говорят о *положительной корре-*

ляционной связи между X и Y, при $r_{xy} < 0$ — об *отрицательной*. Следует учитывать, что коэффициент корреляции характеризует не любую зависимость, а только линейную.

Для нормально распределенной системы двух случайных величин можно доказать, что

$$f(y/x) = \frac{f(x,y)}{f_1(x)} =$$

$$= \frac{1}{\sigma_y \sqrt{1-r^2}\sqrt{2\pi}} \exp\left[-\frac{1}{2(1-r^2)}\left(\frac{y-m_y}{\sigma_y} - r\frac{x-m_x}{\sigma_x}\right)^2\right] =$$

$$= \frac{1}{\sigma_y \sqrt{1-r^2}\sqrt{2\pi}} \exp\left[-\frac{1}{2(1-r^2)\sigma_y^2}\left(y-m_y - r\frac{\sigma_y}{\sigma_x}(x-m_x)\right)^2\right]. \quad (7.25)$$

Условная плотность распределения величины Y соответствует плотности нормального распределения с математическим ожиданием

$$m_{y/x} = m_y + r\frac{\sigma_y}{\sigma_x}(x-m_x) \quad (7.26)$$

и среднеквадратичным отклонением

$$\sigma_{y/x} = \sigma_y \sqrt{1-r^2}\,. \quad (7.27)$$

Величина $m_{y/x}$ называется *условным математическим ожиданием* величины Y при данном X. Линейная зависимость (7.26) — *регрессией Y* на X. По аналогии прямая

$$m_{x/y} = m_x + r\frac{\sigma_x}{\sigma_y}(y-m_y) \quad (7.28)$$

есть регрессия X на Y.

Линии регрессии совпадают только при наличии линейной функциональной зависимости. Из (7.26) и (7.28) видно, что для независимых X и Y линии регрессии параллельны координатным осям.

7.3. Выборочный коэффициент корреляции. Проверка гипотезы об отсутствии корреляции

При обработке результатов большинства физико-химических измерений возникает задача описания зависимости между исследуемыми случайными величинами. Для экспериментального изучения зависимости между двумя случайными величинами X и Y проводят n независимых опытов, при этом в каждом из них получают пару значений

(x_i, y_i), $i = 1, 2, \ldots, n$. О наличии или отсутствии корреляции между X и Y можно качественно судить по виду поля корреляции, нанеся точки (x_i, y_i) на координатную плоскость.

Для количественной оценки тесноты связи служит *выборочный коэффициент корреляции*. Как было установлено ранее, состоятельными и несмещенными оценками для математических ожиданий m_x и m_y служат выборочные средние $\bar{x}$ и $\bar{y}$, а генеральных дисперсий σ_x^2 и σ_y^2 — выборочные дисперсии s_x^2 и s_y^2. Можно доказать, что состоятельной и несмещенной оценкой генеральной ковариации cov_{xy} служит *выборочная ковариация*

$$\mathrm{cov}_{xy}^* = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \qquad (7.29)$$

Пользуясь этой оценкой, рассчитывают выборочный коэффициент корреляции

$$r_{xy}^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, s_x\, s_y}, \qquad (7.30)$$

который является состоятельной оценкой коэффициента корреляции генеральной совокупности со смещением, равным $r(1-r^2)/2n$. Величина смещения убывает с увеличением числа опытов и при $n > 50$ составляет менее 1 %. Выборочный коэффициент корреляции обладает теми же свойствами, что и r_{xy}, и по абсолютной величине также не больше единицы:

$$-1 \le r_{\mathrm{xy}}^* \le 1. \qquad (7.31)$$

Величина выборочного коэффициента корреляции определяет меру криволинейности связи между X и Y. Поэтому возможны случаи, когда при коэффициенте корреляции, значительно меньшем единицы, связь между X и Y оказывается близкой к функциональной, хотя и существенно нелинейной.

В случае, если полученное значение r^* близко к нулю, необходимо провести проверку *гипотезы об отсутствии корреляции* между случайными величинами. Требуется определить, значимо ли отличается r^* от нуля. Если число опытов n достаточно велико (более 20), то в условиях нулевой гипотезы (H_0: $r = 0$) можно использовать нормальное распределение со стандартом

$$\sigma_{r*} \approx (1 - r*^2) / \sqrt{n}. \qquad (7.32)$$

Тогда при β = 0,95 генеральный коэффициент корреляции находится в следующих доверительных границах:

$$r* - \frac{1.96 \cdot (1 - r*^2)}{\sqrt{n}} \le r \le r* + \frac{1.96 \cdot (1 - r*^2)}{\sqrt{n}}. \qquad (7.33)$$

С вероятностью 0,95 можно ожидать, что существует корреляция между случайными величинами, если 0 не содержится внутри доверительного интервала.

На практике, особенно при числе опытов $n < 20$, часто приходится решать вопрос о том, насколько хорошо полученные экспериментальные точки подтверждают линейную связь между величинами X и Y. Ответить на этот вопрос можно следующим образом. Предположим, что две переменные X и Y действительно некоррелированы, т. е. при проведении бесконечно большого числа измерений выборочный коэффициент корреляции для них был бы равен нулю. При конечном числе измерений, однако, маловероятно, чтобы величина r^* была точно равна нулю из-за воздействия случайных факторов.

Обозначим через

$$P_n\left(|r*| \ge r_1*\right)$$

вероятность того, что n измерений двух некоррелированных переменных X и Y приведут к значению $r*$ (по модулю), не меньшему некоторого частного значения r_1*. Результаты расчетов вероятностей P_n для выборок различного объема n и чисел r_1* представлены в табл. 1. Для ответа на вопрос о том, насколько хорошо n пар полученных значений (x_i, y_i) подтверждают линейную связь между исследуемыми величинами, вначале по измеренным точкам вычисляют выборочный коэффициент корреляции r_1*. Далее по табл. 1 находят вероятность P_n того, что n некоррелированных точек приведут к значению коэффициента

Таблица 1

Вероятность P_n того, что n измерений двух некоррелированных переменных дадут коэффициент корреляции $|r^*| \geq r_1^*$ (прочерками отмечены значения, меньшие 0,01)

n	r_1^*								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3	0.94	0.87	0.81	0.74	0.67	0.59	0.51	0.41	0.29
4	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10
5	0.87	0.75	0.62	0.50	0.39	0.28	0.19	0.10	0.04
6	0.85	0.70	0.56	0.43	0.31	0.21	0.12	0.06	0.01
7	0.83	0.67	0.51	0.37	0.25	0.15	0.08	0.03	—
8	0.81	0.63	0.47	0.33	0.21	0.12	0.05	0.02	—
9	0.80	0.61	0.43	0.29	0.17	0.09	0.04	0.01	—
10	0.78	0.58	0.40	0.25	0.14	0.07	0.02	0.01	—
11	0.77	0.56	0.37	0.22	0.12	0.05	0.02	—	—
12	0.76	0.53	0.34	0.20	0.10	0.04	0.01	—	—
13	0.75	0.51	0.32	0.18	0.08	0.03	0.01	—	—
14	0.73	0.49	0.30	0.16	0.07	0.02	0.01	—	—
15	0.72	0.47	0.28	0.14	0.06	0.02	—	—	—
16	0.71	0.46	0.26	0.12	0.05	0.01	—	—	—
17	0.70	0.44	0.21	0.11	0.04	0.01	—	—	—
18	0.69	0.43	0.23	0.10	0.04	0.01	—	—	—
19	0.68	0.41	0.21	0.09	0.03	0.01	—	—	—
20	0.67	0.40	0.20	0.08	0.03	0.01	—	—	—
25	0.63	0.34	0.15	0.05	0.01	—	—	—	—
30	0.60	0.29	0.11	0.03	0.01	—	—	—	—
35	0.57	0.25	0.08	0.02	—	—	—	—	—
40	0.54	0.22	0.06	0.01	—	—	—	—	—
50	0.49	0.16	0.03	—	—	—	—	—	—
60	0.45	0.13	0.02	—	—	—	—	—	—
80	0.38	0.08	0.01	—	—	—	—	—	—
100	0.32	0.05	—	—	—	—	—	—	—

корреляции, не меньшего r_1^*. Если $P_n \leq 0{,}05$ (для «высокозначимых» корреляций $P_n \leq 0{,}01$), то гипотеза о линейной зависимости между величинами X и Y принимается (при выбранном уровне значимости 0,05 или 0,01 соответственно).

Например, по выборке из 5 пар значений (x_i, y_i) получено $r_1^* = 0{,}9$. Вероятность получения коэффициента r^* такого, что $|r^*| \geq 0{,}9$, для 5 некоррелированных точек равна $P_n = 0{,}04$ (табл. 1). Следовательно, гипотеза о линейной связи двух исследуемых величин может быть принята с уровнем значимости 0,05.

7.4. Приближенная регрессия. Метод наименьших квадратов

При исследовании корреляционной зависимости между двумя случайными величинами необходимо по данной выборке объемом n найти *уравнение приближенной регрессии*, чаще всего в виде следующего полинома:

$$\hat{y}(x) = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \ldots = b_0 + \sum_{j=1}^{k} b_j x^j \,, \qquad (7.34)$$

где коэффициенты b_0 и b_j являются оценками соответствующих теоретических коэффициентов истинного уравнения регрессии

$$m_{y/x} = \varphi(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots = \beta_0 + \sum_{j=1}^{k} \beta_j x^j \,, \qquad (7.35)$$

и оценить допускаемую при этом ошибку. Для этого обычно используют метод наименьших квадратов.

Рассмотрим некоторый класс функций, аналитическое выражение которых содержит некоторое число неопределенных коэффициентов, равное k. Наилучшее уравнение приближенной регрессии дает та функция из рассматриваемого класса, для которой сумма квадратов S имеет наименьшее значение:

$$S = \sum_{i=1}^{n} \left[y_i - \hat{y}(x_i) \right]^2 = \min \,. \qquad (7.36)$$

Предположим, что экспериментальные точки отклоняются от уравнения истинной регрессии $\varphi(x)$ только в результате воздействия случайных факторов, а ошибки измерения нормально распределены. Полученные в опытах значения y_i будут распределены по нормальному закону с математическим ожиданием $m(y_i) = \varphi(x_i)$ и дисперсией σ_i^2. При равноточных экспериментах $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2$. Тогда плотность распределения величины Y_i принимает вид

$$f_i(y_i) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{1}{2\sigma^2} \left[y_i - \varphi(x_i) \right]^2 \right\}. \qquad (7.37)$$

В результате опытов случайные величины Y_i приняли совокупность значений y_i. Используем принцип максимального правдоподобия: определим так математические ожидания $\varphi(x_i)$, чтобы вероятность этого события была максимальной. Обозначим через $p_i = f_i(y_i)\,\delta$ вероятность того, что случайная величина Y_i примет значение из ин-

тервала $y_i - \delta/2$, $y_i + \delta/2$. Вероятность совместного осуществления подобных событий для $i = 1, 2, \ldots, n$ равна

$$P = \delta^n \prod_{i=1}^{n} f_i(y_i) = \delta^n \sigma^{-n} (2\pi)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i - \varphi(x_i)\right]^2\right\} =$$

$$= K \exp\left\{-\frac{1}{\sigma^2}\sum_{i=1}^{n}\left[y_i - \varphi(x_i)\right]^2\right\}, \qquad (7.38)$$

где K — коэффициент, не зависящий от $\varphi(x_i)$.

Очевидно, что при заданном σ^2 вероятность P максимальна при условии, что

$$\sum_{i=1}^{n}\left[y_i - \varphi(x_i)\right]^2 = \min.$$

Таким образом, при нормальном распределении случайных величин оптимальность метода наименьших квадратов легко обосновывается.

Нахождение коэффициентов уравнения приближенной регрессии по этому методу связано с задачей определения минимума функции многих переменных. Пусть

$$\hat{y}(x) = f\left(x, b_0, b_1, b_2, \ldots, b_k\right). \qquad (7.40)$$

Требуется найти значения коэффициентов b_0, b_1, b_2, …, b_k так, чтобы

$$S = \sum_{i=1}^{n}\left[y_i - \hat{y}(x_i)\right]^2 = \min.$$

Если S принимает минимальное значение, то

$$\frac{\partial S}{\partial b_0} = 0, \quad \frac{\partial S}{\partial b_1} = 0, \quad \frac{\partial S}{\partial b_2} = 0, \ldots, \quad \frac{\partial S}{\partial b_k} = 0, \qquad (7.41)$$

что соответствует следующей системе уравнений:

$$\sum_{i=1}^{n} 2\left[y_i - \hat{y}(x_i)\right]\frac{\partial \hat{y}(x_i)}{\partial b_0} = 0,$$

$$\sum_{i=1}^{n} 2\left[y_i - \hat{y}(x_i)\right]\frac{\partial \hat{y}(x_i)}{\partial b_1} = 0, \qquad (7.42)$$

…………………………….,

$$\sum_{i=1}^{n} 2\left[y_i - \hat{y}(x_i) \right] \frac{\partial \hat{y}(x_i)}{\partial b_k} = 0 .$$

Преобразуем (7.42)

$$\begin{aligned}
&\sum_{i=1}^{n} y_i \frac{\partial \hat{y}(x_i)}{\partial b_0} - \sum_{i=1}^{n} \hat{y}(x_i) \frac{\partial \hat{y}(x_i)}{\partial b_0} = 0 , \\
&\sum_{i=1}^{n} y_i \frac{\partial \hat{y}(x_i)}{\partial b_1} - \sum_{i=1}^{n} \hat{y}(x_i) \frac{\partial \hat{y}(x_i)}{\partial b_1} = 0 , \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots , \\
&\sum_{i=1}^{n} y_i \frac{\partial \hat{y}(x_i)}{\partial b_k} - \sum_{i=1}^{n} \hat{y}(x_i) \frac{\partial \hat{y}(x_i)}{\partial b_k} = 0 .
\end{aligned} \qquad (7.43)$$

В последней системе содержится столько же $(k+1)$ уравнений, сколько и неизвестных коэффициентов в уравнении (7.40), т. е. она является *системой нормальных уравнений*. Поскольку $S \geq 0$ при любых значениях коэффициентов, то у нее должен существовать по меньшей мере один минимум. Поэтому если система (7.43) имеет единственное решение, то оно и является минимумом для S.

ЛЕКЦИЯ 8

Линейная регрессия от одного параметра. Регрессионный анализ. Аппроксимация, параболическая регрессия. Оценка тесноты нелинейной связи, корреляционный анализ. Метод множественной корреляции.

8.1. Линейная регрессия от одного параметра

Пусть из опытов получена выборка точек (x_i, y_i) объемом n. Найдем методом наименьших квадратов коэффициенты линейного уравнения регрессии

$$\hat{y} = b_0 + b_1 x . \tag{8.1}$$

Система нормальных уравнений уравнений (7.43) с учетом того, что

$$\hat{y}(x_i) = b_0 + b_1 x_i ,$$

принимает вид

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} (b_0 + b_1 x_i) = 0,$$

$$\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} (b_0 + b_1 x_i) x_i = 0, \tag{8.2}$$

или после преобразования

$$n b_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i ,$$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i . \tag{8.3}$$

Решив систему уравнений, получим

$$b_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}, \tag{8.4}$$

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} =$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, s_x^2}. \qquad (8.5)$$

Из системы уравнений (8.3) видно, что между коэффициентами b_0 и b_1 существует корреляционная зависимость, выражение для которой можно получить, например, из первого уравнения системы:

$$b_0 = \bar{y} - b_1 \bar{x}. \qquad (8.6)$$

Выборочный коэффициент корреляции с учетом (8.5) равен

$$r_{xy}^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, s_x\, s_y} = \frac{b_1 (n-1) s_x^2}{(n-1)\, s_x\, s_y} = \frac{b_1 s_x}{s_y} \qquad (8.7)$$

и оценивает силу линейной связи между Y и X.

8.2. Регрессионный анализ

Итак, уравнение линейной регрессии определено. Проведем статистический анализ полученных результатов, заключающийся в оценке значимости коэффициентов регрессии и проверки адекватности полученного уравнения экспериментальным данным. Подобный анализ и называется *регрессионным*.

Примем, что

1) входной параметр x измеряется с гораздо большей точностью по сравнению с выходной величиной y;
2) значения y_i получены независимым образом и нормально распределены;
3) если при каждом заданном значении x_i проводится серия параллельных опытов, то выборочные дисперсии s_i^2 однородны.

8.2.1. Проверка адекватности приближенного уравнения регрессии эксперименту

Рассмотрим три наиболее часто встречающихся варианта проверки адекватности полученного уравнения регрессии.

1. Пусть при каждом значении x_i проведена серия из m параллельных опытов. Тогда дисперсия воспроизводимости с числом степеней свободы $f_{\text{воспр.}} = n\,(m-1)$ равна

$$s^2_{\text{воспр.}} = \frac{\sum_{i=1}^{n} s_i^2}{n}. \qquad (8.8)$$

Дисперсия адекватности определяется формулой

$$s^2_{\text{ад.}} = \frac{m\sum_{i=1}^{n}\left(\bar{y}_i - \hat{y}(x_i)\right)^2}{n-l}, \qquad (8.9)$$

где l — число коэффициентов в уравнении регрессии (при линейной регрессии $l = 2$),

$$\bar{y}_i = \frac{1}{m}\sum_{u=1}^{m} y_{iu}. \qquad (8.10)$$

Число степеней свободы дисперсии адекватности равно $f_{\text{ад.}} = n - l$.

Адекватность уравнения проверяется по критерию Фишера

$$F = s^2_{\text{ад.}} / s^2_{\text{воспр.}}. \qquad (8.11)$$

Если вычисленное значение F окажется меньше табличной величины $F_{1-p}(f_1, f_2)$ для уровня значимости p и числа степеней свободы $f_1 = f_{\text{ад.}}$ и $f_2 = f_{\text{воспр.}}$, то уравнение адекватно эксперименту.

2. Основная серия опытов проведена без параллельных, а дисперсия воспроизводимости определена в отдельной серии из m опытов, тогда

$$s^2_{\text{ад.}} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}(x_i)\right)^2}{n-l}, \qquad (8.12)$$

$$s^2_{\text{воспр.}} = \frac{\sum\limits_{u=1}^{m}\left(y_u^0 - \overline{y^0}\right)^2}{m-1}, \quad \text{где } \overline{y^0} = \frac{1}{m}\sum_{u=1}^{m} y_i^0 . \tag{8.13}$$

Адекватность уравнения проверяется по критерию Фишера (8.11), при этом $f_2 = f_{\text{воспр.}} = m - 1$.

3. Основная серия опытов выполнена без параллельных, и нет данных для расчета дисперсии воспроизводимости. Тогда по критерию Фишера сравнивается дисперсия адекватности и дисперсия относительно среднего

$$F = \frac{s_y^2(f_1)}{s^2_{\text{ад.}}(f_2)}, \tag{8.14}$$

где

$$s_y^2 = \frac{\sum\limits_{i=1}^{n}\left(y_i - \overline{y}\right)^2}{n-1}. \tag{8.15}$$

Чем больше полученное F превышает табличное $F_{1\text{-}p}(f_1, f_2)$ для уровня значимости p и чисел степеней свободы $f_1 = n - 1$ и $f_2 = n - l$, тем эффективнее уравнение регрессии.

8.2.2. Оценка значимости коэффициентов уравнения регрессии

Значимость коэффициентов уравнения регрессии оценивается по критерию Стъюдента

$$t_j = \frac{\left|b_j\right|}{s(b_j)}, \tag{8.16}$$

где b_j — j-й коэффициент уравнения регрессии; $s(b_j)$ — среднее квадратичное отклонение j-го коэффициента. Если t_j больше табличной величины $t_{1\text{-}p/2}$ для выбранного уровня значимости p и числа степеней свободы f дисперсии j-го коэффициента, то коэффициент b_j значимо отличается от нуля.

В случае линейной регрессии средние квадратичные отклонения коэффициентов рассчитываются следующим образом:

$$s(b_0) = \sqrt{\left(s^2 \sum_{i=1}^{n} x_i^2\right) \Bigg/ \left(n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)}, \qquad (8.17)$$

$$s(b_1) = \sqrt{\left(s^2\, n\right) \Bigg/ \left(n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)}, \qquad (8.18)$$

где дисперсия s^2 в общем случае определяется как

$$s^2 = \frac{f_{\text{воспр.}} s^2_{\text{воспр.}} + f_{\text{ад.}} s^2_{\text{ад.}}}{f_{\text{воспр.}} + f_{\text{ад.}}} = \frac{n(m-1)\, s^2_{\text{воспр.}} + (n-l)\, s^2_{\text{ад.}}}{n(m-1) + (n-l)}. \qquad (8.19)$$

Число степеней свободы средневзвешенной дисперсии s^2 равно

$$f = n(m-1) + (n-l) = nm - n + n - l = nm - l.$$

Дисперсии воспроизводимости и адекватности рассчитываются по формулам (8.8) и (8.9) или (8.12) и (8.13). Если у экспериментатора нет оснований сомневаться в линейном характере изучаемой зависимости и опыты проведены без параллельных (т. е. $m = 1$), то $s^2 = s^2_{\text{ад.}}$ и $f = f_{\text{ад.}} = n - l$. Дисперсия адекватности в этом случае определяется по формуле (8.12).

Для оценки случайных ошибок в определении коэффициентов приближенного уравнения регрессии можно также воспользоваться критерием Стъюдента. Рассмотрим величину

$$t = \frac{b_0 - \beta_0}{s(b_0)}, \qquad (8.20)$$

где β_0 — истинное значение коэффициента b_0. Произведя выкладки, аналогичные представленным в лекции 4, получим

$$b_0 - s(b_0)\, t_{1-p/2} \le \beta_0 \le b_0 + s(b_0)\, t_{1-p/2}, \qquad (8.21)$$

или

$$\beta_0 = b_0 \pm s(b_0)\, t_{1-p/2}, \qquad (8.22)$$

где $t_{1\text{-}p/2}$ — квантиль t-распределения для числа степеней свободы f и выбранного уровня значимости p.

Аналогично можно построить доверительный интервал для коэффициента b_1:

$$b_1 - s(b_1)\, t_{1-p/2} \le \beta_1 \le b_1 + s(b_1)\, t_{1-p/2}, \qquad (8.23)$$

$$\beta_1 = b_1 \pm s(b_1)\, t_{1-p/2}\,. \qquad (8.24)$$

С учетом (8.22) и (8.24), уравнение регрессии принимает следующий вид:

$$y = \beta_0 + \beta_1 x = (b_0 \pm s(b_0)\, t_{1-p/2}) + (b_1 \pm s(b_1)\, t_{1-p/2})\, x\,.$$

8.2.3. Оценка доверительного интервала для искомой функции

На практике нередко возникает необходимость в оценке точек, резко выделяющихся из общей линейной закономерности. Подобную оценку легко произвести, построив доверительный интервал («*коридор ошибок*») искомой функции. Под «*коридором ошибок*» понимают границы, отсчитываемые по обе стороны от полученной прямой и показывающие пределы, в которых должны лежать экспериментальные точки. Точки, лежащие за пределами этого коридора, следует признать ошибочными и исключить из общей выборки.

Воспользуемся критерием Стъюдента и рассмотрим величину

$$t = \frac{\hat{y} - m_{y/x}}{s(\hat{y})}, \qquad (8.25)$$

где $m_{y/x}$ — условное математическое ожидание Y при заданном X; $s(\hat{y})$ — выборочное среднеквадратичное отклонение, соответствующее выборочной дисперсии

$$s^2(\hat{y}) = s^2(b_0) + \left(x^2 - 2x\bar{x}\right) s^2(b_1) \qquad (8.26)$$

с числом степеней свободы $f = nm - 2$, если среднеквадратичные отклонения коэффициентов рассчитываются на основе средневзвешенной дисперсии s^2, определяемой по формуле (8.19), и $f = n - 2$, если $s^2 = s^2_{\text{ад.}}$. Тогда границы коридора ошибок для произвольного значения аргумента x определяются следующим выражением:

$$m_{y/x} = \hat{y}(x) \pm t_{1-p/2} \cdot s(\hat{y}), \qquad (8.27)$$

где $t_{1\text{-}p/2}$ — квантиль t-распределения для числа степеней свободы f и выбранного уровня значимости p (обычно 0,05).

Процедура выделения из общей совокупности точек, содержащих грубые ошибки, заключается в следующем. Вначале методом наименьших квадратов обрабатываются все полученные экспериментальные данные, не выбрасывая ни одной точки. Далее по формуле (8.27) для каждой ординаты (для каждого заданного значения x) определяет-

ся доверительный интервал при выбранной доверительной вероятности. *Если оказывается, что одна или несколько точек при этом выпадают из рассчитанных для них интервалов и величина отклонения превышает систематическую погрешность измерения, то их следует признать ошибочными и исключить из рассмотрения.* Затем весь расчет коэффициентов, их случайных ошибок и коридора ошибок повторяется заново.

8.3. Оценка тесноты нелинейной связи

Если уравнение регрессии получено с достаточной точностью, то силу стохастической связи между величинами Y и X можно охарактеризовать величиной

$$\gamma = \frac{(n-l)\,s_{\text{ад.}}^2}{(n-1)\,s_y^2}. \tag{8.28}$$

Дисперсия адекватности (остаточная дисперсия) и дисперсия относительно среднего рассчитываются по формулам (8.12) и (8.15) соответственно. Связь тем сильнее, чем меньше γ. Величина

$$\Theta = \sqrt{1-\gamma} \tag{8.29}$$

называется *корреляционным отношением*, для которого справедливо

$$0 \le \Theta \le 1. \tag{8.30}$$

Чем больше Θ, тем сильнее связь.

В общем случае анализ силы связи по корреляционному отношению называют *корреляционным анализом*. Функциональная зависимость между случайными величинами существует, если $\Theta = 1$. Однако при $\Theta = 0$ однозначно говорить об отсутствии связи можно только в случае нормального распределения случайных величин.

При линейной регрессии корреляционное отношение равно коэффициенту корреляции:

$$\Theta = \sqrt{1-\frac{(n-2)\,s_{\text{ад.}}^2}{(n-1)\,s_y^2}} = |r^*|. \tag{8.31}$$

8.4. Аппроксимация. Параболическая регрессия

В общем случае при описании функциональной зависимости между двумя случайными величинами используют полиномы некоторой степени, коэффициенты которых могут и не иметь определенного физического смысла. Такая операция называется *аппроксимацией* экспериментальных данных. Полученная эмпирическая формула обычно справедлива только для сравнительно узкого интервала измерений и неприменима вне этого интервала. При использовании метода наименьших квадратов коэффициенты приближенного уравнения регрессии определяются решением системы линейных уравнений.

Допустим, что зависимость между величинами X и Y описывается параболой второго порядка

$$\hat{y}(x) = b_0 + b_1 x + b_2 x^2. \qquad (8.32)$$

Тогда

$$\frac{\partial \hat{y}(x)}{\partial b_0} = 1, \quad \frac{\partial \hat{y}(x)}{\partial b_1} = x, \quad \frac{\partial \hat{y}(x)}{\partial b_2} = x^2, \qquad (8.33)$$

и система нормальных уравнений (7.43) принимает вид

$$b_0 n + b_1 \sum_{i=1}^{n} x_i + b_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i,$$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 + b_2 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i, \qquad (8.34)$$

$$b_0 \sum_{i=1}^{n} x_i^2 + b_1 \sum_{i=1}^{n} x_i^3 + b_2 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i.$$

Решая систему (8.34), находят коэффициенты искомой квадратичной функции. При описании функциональных зависимостей полиномами большей степени коэффициенты определяются из аналогичных по структуре систем уравнений.

На практике адекватности уравнения регрессии эксперименту добиваются повышением степени аппроксимирующего полинома. При использовании полинома k-степени требуется определять $k+1$ коэффициент. Увеличение степени полинома прекращают, если дисперсия адекватности (остаточная дисперсия) уравнения регрессии $k+1$ степени (s_{k+1}^2) перестает быть значимо меньше дисперсии адекватности,

вычисленной для полинома k-степени (s_k^2). Значимость различия исследуется по критерию Фишера

$$F = s_k^2 / s_{k+1}^2,$$

где

$$s_k^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}(x_i))^2}{n-(k+1)}, \quad s_{k+1}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}(x_i))^2}{n-(k+2)}. \qquad (8.35)$$

Если полученное F меньше табличного $F_{1-p}(f_1, f_2)$ для уровня значимости p и чисел степеней свободы $f_1 = f_k = n - k - 1$ и $f_2 = f_{k+1} = n - k - 2$, то увеличение степени полинома нужно прекратить и в качестве приближенного уравнения регрессии использовать полином k-степени.

8.5. Приведение некоторых функциональных зависимостей к линейному виду

При малых объемах выборки увеличение порядка полинома может иногда приводить к росту остаточной дисперсии. Чтобы избежать этого, при решении многих задач производят замену переменных. Например, зависимости типа

$$\hat{z} = a_0\, a_1^x \text{ или } \hat{z} = a_0\, t^{a_1} \qquad (8.36)$$

сводятся к линейным $\hat{y} = b_0 + b_1 x$ следующим образом:

$$\ln\hat{z} = \ln a_0 + x\ln a_1, \quad \hat{y} = \ln\hat{z}, b_0 = \ln a_0, b_1 = \ln a_1, \qquad (8.37)$$

$$\ln\hat{z} = \ln a_0 + a_1 \ln t, \quad \hat{y} = \ln\hat{z}, b_0 = \ln a_0, b_1 = a_1, x = \ln t. \qquad (8.38)$$

Коэффициенты уравнений (8.37) и (8.38) находятся методом наименьших квадратов.

Рассмотрим некоторые наиболее часто встречающиеся случаи линеаризации зависимостей при обработке результатов физико-химических экспериментов.

1. Температурная зависимость контанты равновесия реакции для небольшого интервала температур имеет вид

$$\ln K = \frac{\Delta S}{R} - \frac{\Delta H}{R} \cdot \frac{1}{T}, \qquad (8.39)$$

где ΔS и ΔH — энтропия и энтальпия реакции. Непосредственно измеряемыми величинами являются константа равновесия K и температура T. Произведем замену переменных:

$$\hat{y} = b_0 + b_1 x, \text{ где } \hat{y} = \ln K,\ b_0 = \frac{\Delta S}{R},\ b_1 = -\frac{\Delta H}{R},\ x = \frac{1}{T}.$$

Коэффициенты b_0 и b_1 определяются методом наименьших квадратов.

Энтальпия и энтропия реакции с учетом случайных ошибок равны

$$\Delta S = R \cdot (b_0 \pm s(b_0)\, t_{1-p/2}) = R b_0 \pm R s(b_0)\, t_{1-p/2},$$

$$\Delta H = -R \cdot (b_1 \pm s(b_1)\, t_{1-p/2}) = -(R b_1 \pm R s(b_1)\, t_{1-p/2}).$$

2. Температурная зависимость давления насыщенного пара вещества в узком интервале температур имеет вид

$$\ln P = a - \frac{\Delta H}{R} \cdot \frac{1}{T}, \tag{8.40}$$

где a — константа, ΔH —энтальпия парообразования (испарения или сублимации). Непосредственно определяемыми величинами являются давление насыщенного пара P и температура T. Произведем замену переменных:

$$\hat{y} = b_0 + b_1 x, \text{ где } \hat{y} = \ln P,\ a = b_0,\quad b_1 = -\frac{\Delta H}{R},\ x = \frac{1}{T}.$$

Энтальпия парообразования с учетом случайной ошибки равна

$$\Delta H = -R \cdot (b_1 \pm s(b_1) t_{1-p/2}) = -(R b_1 \pm R s(b_1)\, t_{1-p/2}).$$

3. Константа скорости реакции первого порядка описывается следующим уравнением:

$$k = \frac{1}{t} \ln \frac{C_0}{C}, \tag{8.41}$$

или

$$\ln C = \ln C_0 - k t,$$

где k — константа скорости реакции, C_0 и C — исходная и текущая концентрация реагирующего вещества к моменту времени t соответственно. Произведем замену переменных:

$$\hat{y} = b_0 + b_1 x, \text{ где } \hat{y} = \ln C, b_0 = \ln C_0, b_1 = k, x = t.$$

Определив коэффициент b_1 методом наименьших квадратов, получим значение константы скорости реакции с учетом случайной ошибки:

$$k = -(b_1 \pm s(b_1)\, t_{1-p/2}).$$

8.6. Метод множественной корреляции

На практике часто бывает необходимым исследовать корреляционную связь между многими (а не только двумя) величинами. В случае, когда необходимо установить зависимость величины Y от более чем одного параметра, обычно используют уравнения множественной регрессии следующего вида

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k. \tag{8.42}$$

Коэффициенты уравнения находят методом наименьших квадратов, т. е. определяют из условия

$$S = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \min, \tag{8.43}$$

где $\hat{y}_i = \hat{y}(x_{1i}, x_{2i}, \ldots, x_{ki})$. Условия минимума функции S следующие:

$$\frac{\partial S}{\partial b_0} = 0, \quad \frac{\partial S}{\partial b_1} = 0, \quad \ldots \quad , \quad \frac{\partial S}{\partial b_k} = 0. \tag{8.44}$$

Коэффициенты уравнения приближенной регрессии находят из решения системы $(k+1)$ нормальных уравнений, полученных из условий (8.44).

Рассмотрим случай, когда величина Y линейно зависит от двух переменных X_1 и X_2. Пусть из опытов получена выборка точек (x_{1i}, x_{2i}, y_i) объемом n. Найдем методом наименьших квадратов коэффициенты линейного уравнения регрессии

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2. \tag{8.45}$$

Тогда

$$\frac{\partial \hat{y}}{\partial b_0} = 1, \quad \frac{\partial \hat{y}}{\partial b_1} = x_1, \quad \frac{\partial \hat{y}}{\partial b_2} = x_2. \tag{8.46}$$

Система нормальных уравнений, соответствующих условиям (8.44), принимает следующий вид:

$$\sum_{i=1}^{n} 2\left[y_i - \hat{y}_i \right] \frac{\partial \hat{y}}{\partial b_0} = 0,$$

$$\sum_{i=1}^{n} 2\left[y_i - \hat{y}_i\right]\frac{\partial \hat{y}}{\partial b_1} = 0, \qquad (8.47)$$

$$\sum_{i=1}^{n} 2\left[y_i - \hat{y}_i\right]\frac{\partial \hat{y}}{\partial b_2} = 0.$$

С учетом того, что $\hat{y}_i = b_0 + b_1 x_{1_i} + b_2 x_{2_i}$ и значений частных производных (8.46), после арифметических преобразований получаем

$$b_0 n + b_1\sum_{i=1}^{n} x_{1_i} + b_2\sum_{i=1}^{n} x_{2_i} = \sum_{i=1}^{n} y_i,$$

$$b_0\sum_{i=1}^{n} x_{1_i} + b_1\sum_{i=1}^{n} x_{1_i}^2 + b_2\sum_{i=1}^{n} x_{1_i}x_{2_i} = \sum_{i=1}^{n} x_{1_i}y_i, \qquad (8.48)$$

$$b_0\sum_{i=1}^{n} x_{2_i} + b_1\sum_{i=1}^{n} x_{1_i}x_{2_i} + b_2\sum_{i=1}^{n} x_{2_i}^2 = \sum_{i=1}^{n} x_{2_i}y_i.$$

Решая полученную систему уравнений относительно b_0, b_1 и b_2, находим наилучшую аппроксимацию для соотношения (8.45). Силу линейной связи между переменными X_1 и X_2 можно оценить на основании выборочного коэффициента корреляции

$$r^*(x_1, x_2) = \frac{\sum_{i=1}^{n}\left(x_{1_i} - \overline{x_1}\right)\left(x_{2_i} - \overline{x_2}\right)}{(n-1)\cdot s(x_1)\cdot s(x_2)}. \qquad (8.49)$$

ЛЕКЦИЯ 9

Дисперсионный анализ, его задачи. Проведение однофакторного и двухфакторного дисперсионного анализа.

9.1. Задачи дисперсионного анализа. Однофакторный дисперсионный анализ

Средние значения измеряемых величин зависят от комплекса основных факторов (качественных и количественных), определяющих условия проведения опыта, и случайных факторов. *Задачей дисперсионного анализа и является изучение влияния тех или иных факторов на изменчивость средних.* В зависимости от числа источников дисперсии (числа рассматриваемых факторов) различают однофакторный и многофакторный дисперсионный анализ. Многофакторный дисперсионный анализ более эффективен по сравнению с классическим методом исследования, при котором изменяется только один фактор при постоянстве всех остальных, что не позволяет определить влияние взаимодействия различных факторов на результаты эксперимента.

При дисперсионном анализе каждое наблюдение используется для одновременной оценки всех факторов и их взаимодействий. Суть дисперсионного анализа заключается в выделении и оценке отдельных факторов, влияющих на значения среднего. При этом суммарная выборочная дисперсия разлагается на составляющие, обусловленные действием независимых факторов. Влияние данного фактора признается значимым, если соответствующая ему выборочная дисперсия значимо отличается от дисперсии воспроизводимости, обусловленной случайными ошибками. *Проверка значимости оценок дисперсий проводится по критерию Фишера.*

В дальнейшем примем, что:

1) случайные ошибки нормально распределены;
2) эксперименты равноточны;
3) изучаемые факторы влияют только на изменчивость средних, но не на дисперсию наблюдений (она постоянна).

При дисперсионном анализе рассматриваются факторы двух видов: *со случайными уровнями* и с *фиксированными*. В первом случае выбор уровней фактора производится из бесконечной совокупности возможных значений. Если все уровни выбираются случайным образом, то математическая модель объекта называется *моделью со случайными уровнями факторов*. Если же каждый фактор может принимать только некоторые из фиксированных значений, то говорят о *модели с фиксированными уровнями факторов*. В случае *модели сме-*

шанного типа одна группа факторов рассматривается на случайных уровнях, а другая — на фиксированных.

Рассмотрим влияние на результаты опытов единичного фактора A, принимающего k различных значений (фактор A имеет k фиксированных уровней a_i, $i = 1, 2, \ldots, k$). Обозначим через y_{ij} результат j-опыта в серии из n_i числа измерений ($j = 1, 2, \ldots, n_i$), выполненных на i-уровне фактора A (табл. 2).

Таблица 2

Исходные данные для однофакторного дисперсионного анализа

Номер наблюдения	Уровни фактора A			
	a_1	a_2	…	a_k
1	y_{11}	y_{21}	…	y_{k1}
2	y_{12}	y_{22}	…	y_{k2}
…	…	…	…	…
n	y_{1n}	y_{2n}	…	y_{kn}
Итоги:	B_1, C_1	B_2, C_2	…	B_k, C_k

Предположим, что результат каждого опыта можно представить в виде следующей модели:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad (9.1)$$

где μ — суммарный эффект во всех опытах; α_i — эффект, обусловленный влиянием фактора A на i-уровне; ε_{ij} — случайная ошибка опыта на i-уровне. Примем также, что наблюдения на фиксированном уровне фактора A нормально распределены относительно среднего значения $(\mu + \alpha_i)$ с общей дисперсией $\sigma^2_{\text{ош.}}$. Для того чтобы решить вопрос о значимости влияния фактора A, следует проверить нулевую гипотезу равенства математических ожиданий сумм $(\mu + \alpha_i)$ на различных уровнях этого фактора:

$$H_0 : m_1 = m_2 = \ldots = m_k = m, \qquad (9.2)$$

где $m_i = M\{\mu + \alpha_i\}$.

Рассмотрим случай, когда на каждом уровне выполнено равное число опытов ($n_1 = n_2 = \ldots = n_k = n$). Общее число опытов равно

$$N = n_1 + n_2 + \ldots + n_k = kn. \qquad (9.3)$$

Обозначим сумму результатов всех опытов (итогов) на i-уровне через

$$B_i = \sum_{j=1}^{n} y_{ij}\,, \tag{9.4}$$

а сумму квадратов итогов на i-уровне через

$$C_i = \sum_{j=1}^{n} {y_{ij}}^2\,. \tag{9.5}$$

Тогда среднее значение наблюдений на i-уровне равно

$$\overline{y_i} = \frac{\sum\limits_{j=1}^{n} y_{ij}}{n} = \frac{B_i}{n}\,, \tag{9.6}$$

а общее среднее для всей выборки из N наблюдений —

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} = \frac{1}{k}\sum_{i=1}^{k}\overline{y_i} = \frac{1}{kn}\sum_{i=1}^{k} B_i\,. \tag{9.7}$$

Общая выборочная дисперсия опытов определяется выражением

$$s^2 = \frac{\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n}(y_{ij}-\overline{y})^2}{N-1} = \frac{1}{N-1}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}{y_{ij}}^2 - \frac{1}{N}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}\right)^2\right] =$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{k} C_i - \frac{1}{N}\left(\sum_{i=1}^{k} B_i\right)^2\right], \tag{9.8}$$

а выборочная дисперсия на i-уровне —

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_{ij}-\overline{y_i})^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n}{y_{ij}}^2 - \frac{1}{n}\left(\sum_{j=1}^{n} y_{ij}\right)^2\right] =$$

$$= \frac{1}{n-1}\left[C_i - \frac{B_i^2}{n}\right]. \tag{9.9}$$

Если выборочные дисперсии s_i^2 однородны (проверка по критерию Кохрена), то лучшей оценкой дисперсии $\sigma^2_{\text{ош.}}$, характеризующей влияние случайных факторов, будет выборочная дисперсия

$$s_{\text{ош}}^2 = \frac{1}{k}\sum_{i=1}^{k} s_i^2 \qquad (9.10)$$

с числом степеней свободы $f_{\text{ош}} = k(n-1) = N - k$. Приближенно оценить дисперсию фактора A можно следующим образом:

$$\sigma_A^2 \approx s^2 - s_{\text{ош}}^2. \qquad (9.11)$$

Для получения более точной оценки рассмотрим отклонение средних на фиксированных уровнях от общего среднего:

$$\frac{1}{k-1}\sum_{i=1}^{k} (\bar{y}_i - \bar{y})^2 \approx \sigma_A^2 + \frac{\sigma_{\text{ош}}^2}{n} \approx \sigma_A^2 + \frac{s_{\text{ош}}^2}{n}. \qquad (9.12)$$

В данном случае под дисперсией фактора A понимают математическое ожидание среднего квадрата отклонений, обусловленного влиянием этого фактора. Выборочная дисперсия

$$s_A^2 = \frac{n}{k-1}\sum_{i=1}^{k} (\bar{y}_i - \bar{y})^2 \approx n\sigma_A^2 + s_{\text{ош}}^2 \qquad (9.13)$$

с числом степеней свободы $f_A = k - 1$ используется для проверки нулевой гипотезы (9.2) по критерию Фишера.

При этом, если нулевая гипотеза ($H_0 : \sigma_{\text{A}}^2 = \sigma_{\text{ош}}^2$) верна, выполняется следующее условие:

$$\left(s_A^2 \,/\, s_{\text{ош}}^2\right) \le F_{1-p}, \qquad (9.14)$$

т. е. различие между дисперсиями s_A^2 и $s_{\text{ош}}^2$ является незначимым, и следовательно влияние фактора A на результаты опытов тоже незначимо (сопоставимо с эффектом случайности). При проверке гипотезы используется односторонний критерий, так как альтернативной гипотезой является $H_1 : \sigma_A^2 > \sigma_{\text{ош}}^2$. Если же

$$\left(s_A^2 \,/\, s_{\text{ош}}^2\right) > F_{1-p}, \qquad (9.15)$$

то нулевая гипотеза о равенстве математических ожиданий сумм $(\mu + \alpha_i)$ отвергается (влияние фактора A значимо). Чтобы выяснить, какие средние различны, можно использовать критерий Стъюдента, сравнивая средние попарно. Оценить влияние фактора A можно на основании (9.13):

$$\sigma_A^2 = \frac{s_A^2 - s_{\text{ош}}^2}{n}. \qquad (9.16)$$

Если на каждом уровне выполнено разное число опытов, выборочная дисперсия фактора A рассчитывается по формуле

$$s_A^2 = \frac{1}{k-1}\left[\sum_{i=1}^{k}\frac{B_i^2}{n_i} - \frac{1}{N}\left(\sum_{i=1}^{k}B_i\right)^2\right], \tag{9.17}$$

а выборочная дисперсия, характеризующая влияние случайных факторов, по формуле

$$s_{\text{ош}}^2 = \sum_{i=1}^{k} f_i s_i^2 \bigg/ \sum_{i=1}^{k} f_i, \tag{9.18}$$

где $f_i = n_i - 1$. Число степеней свободы $s_{\text{ош}}^2$ равно $f_{\text{ош}} = N - k$.

Если дисперсия s_A^2 значимо отличается от дисперсии $s_{\text{ош}}^2$, т. е. выполняется неравенство (9.15), то дисперсия фактора A оценивается по формуле

$$\sigma_A^2 \approx \frac{k-1}{N-1}\left(s_A^2 - s_{\text{ош}}^2\right). \tag{9.19}$$

9.2. Двухфакторный дисперсионный анализ

Рассмотрим влияние на результаты опытов двух факторов A и B. Фактор A исследуется на k уровнях ($i = 1, 2, \ldots, k$), фактор B — на m уровнях ($j = 1, 2, \ldots, m$). Пусть при каждом сочетании уровней факторов выполнено n параллельных опытов ($q = 1, 2, \ldots, n$). Тогда общее число опытов равно $N = nkm$. Обозначим через y_{ijq} результат q-го опыта, выполненного на i-уровне фактора A и j-уровне фактора B.

Предположим, что результат каждого опыта можно представить следующим образом:

$$y_{ijq} = \mu + \alpha_i + \beta_j + \alpha_i\beta_j + \varepsilon_{ijq}, \tag{9.20}$$

где μ — общее среднее (суммарный эффект во всех опытах); α_i и β_j — эффекты, обусловленные влиянием фактора A на i-уровне и фактором B на j-уровне соответственно; ε_{ijq} — случайная ошибка опыта, распределенная нормально с нулевым математическим ожиданием и дисперсией $\sigma_{\text{ош}}^2$; $\alpha_i\beta_j$ — эффект взаимодействия факторов. Величина $\alpha_i\beta_j$ характеризует отклонение среднего в (ij)-серии опытов от суммы первых трех членов в ур-и (9.20), а соответствующую ей дисперсию σ_{AB}^2 можно оценить только при наличии параллельных опытов.

При отсутствии параллельных опытов (табл. 3) или в случае, если эффектом взаимодействия факторов пренебрегают, для описания результатов экспериментов используется линейная модель

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}. \qquad (9.21)$$

Таблица 3

Исходные данные для двухфакторного дисперсионного анализа без параллельных опытов. Факторы *A* и *B* исследуются на 3 уровнях

Уровни фактора *B*	Уровни фактора *A*			
	a_1	a_2	a_3 (a_k)	Средние:
b_1	y_{11}	y_{21}	y_{31} (y_{k1})	$\bar{y}'_1$
b_2	y_{12}	y_{22}	y_{32} (y_{k2})	$\bar{y}'_2$
b_3 (b_m)	y_{13}	y_{23}	y_{33} (y_{km})	$\bar{y}'_3(\bar{y}'_m)$
Средние:	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3(\bar{y}_k)$	—

Обозначим через $\bar{y}_i$ и $\bar{y}'_j$ средние по столбцам и по строкам:

$$\bar{y}_i = \frac{\sum_{j=1}^{m} y_{ij}}{m}, \quad \bar{y}'_j = \frac{\sum_{i=1}^{k} y_{ij}}{k}, \qquad (9.22)$$

а через $\bar{y}$ — среднее всех опытов:

$$\bar{y} = \frac{1}{km}\sum_{i=1}^{k}\sum_{j=1}^{m} y_{ij} = \frac{1}{k}\sum_{i=1}^{k} \bar{y}_i. \qquad (9.23)$$

Рассмотрим влияние факторов *A* и *B* на рассеяние средних по столбцам и по строкам соответственно относительно общего среднего. Рассеяние в средних по строкам не зависит от фактора *A*, так как все его уровни усреднены, и определяется влиянием фактора *B* и случайных факторов.

Тогда с учетом того, что дисперсия среднего в *k* раз меньше дисперсии случайной ошибки единичного измерения, имеем

$$\sigma_B^2 + \frac{\sigma_{\text{ош}}^2}{k} \approx \frac{1}{m-1}\sum_{j=1}^{m}\left(\bar{y}'_j - \bar{y}\right)^2. \qquad (9.24)$$

Аналогичным образом можно показать, что

$$\sigma_A^2 + \frac{\sigma_{\text{ош}}^2}{m} \approx \frac{1}{k-1}\sum_{i=1}^{k}\left(\bar{y}_i - \bar{y}\right)^2 . \tag{9.25}$$

Таким образом, чтобы оценить дисперсии факторов A и B, необходимо знать дисперсию случайной ошибки.

Оценить влияние случайных факторов при отсутствии параллельных опытов можно следующим образом. Рассеяние результатов опытов в i-столбце относительно его среднего обусловлено влиянием фактора B и фактора случайности:

$$s_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}\left(y_{ij} - \bar{y}_i\right)^2 \approx \sigma_B^2 + \sigma_{\text{ош}}^2 . \tag{9.26}$$

Равенство (9.26) станет более точным, если использовать средневзвешенное значение дисперсии по всем столбцам:

$$\sigma_B^2 + \sigma_{\text{ош}}^2 \approx \frac{1}{k}\sum_{i=1}^{k} s_i^2 . \tag{9.27}$$

Вычитая (9.24) из (9.27), получим

$$\sigma_{\text{ош}}^2 - \frac{\sigma_{\text{ош}}^2}{k} \approx \frac{1}{k}\sum_{i=1}^{k} s_i^2 - \frac{1}{m-1}\sum_{j=1}^{m}\left(\bar{y}'_j - \bar{y}\right)^2 , \tag{9.28}$$

или после арифметических преобразований

$$\sigma_{\text{ош}}^2 \approx \frac{1}{(k-1)(m-1)}\left[(m-1)\sum_{i=1}^{k} s_i^2 - k\sum_{j=1}^{m}\left(\bar{y}'_j - \bar{y}\right)^2\right] \cong s_{\text{ош}}^2 . \tag{9.29}$$

Полученную оценку для дисперсии случайной ошибки с числом степеней свободы $f_{\text{ош}} = (k-1)(m-1)$ обозначим через $s_{\text{ош}}^2$. Определим также следующие выборочные дисперсии:

$$s_A^2 = \frac{m}{k-1}\sum_{i=1}^{k}\left(\bar{y}_i - \bar{y}\right)^2 \approx m\sigma_A^2 + s_{\text{ош}}^2 , \tag{9.30}$$

$$s_B^2 = \frac{k}{m-1}\sum_{j=1}^{m}\left(\bar{y}'_j - \bar{y}\right)^2 \approx k\sigma_B^2 + s_{\text{ош}}^2 \tag{9.31}$$

с числом степеней свободы $f_A = (k-1)$ и $f_B = (m-1)$.

Проверка нулевой гипотезы о незначимости влияния факторов A и B проводится по критерию Фишера: если

$$\frac{s_A^2}{s_{\text{ош}}^2} \le F_{1-p}(f_A, f_{\text{ош}}) \text{ и (или) } \frac{s_B^2}{s_{\text{ош}}^2} \le F_{1-p}(f_B, f_{\text{ош}}), \tag{9.32}$$

то влияние фактора признается незначимым ($\alpha_i = 0$ и (или) $\beta_j = 0$).

Если одно (или оба) из неравенств (9.32) не выполняется, то влияние соответствующего фактора (факторов) значимо. Определить, какие именно средние различны, можно по критерию Стъюдента.

Рассмотрим теперь случай, когда при каждом сочетании уровней факторов A и B выполнено n параллельных опытов ($u = 1, 2, \ldots, n$), что дает возможность оценить влияние взаимодействия этих факторов на результаты опытов.

Так, например, в табл. 3 вместо одного значения y_{11} появится серия значений $y_{111}, y_{112}, \ldots, y_{11n}$. Обозначим через $\bar{y}_{ij}$ среднее в ячейке (среднее серии параллельных опытов):

$$\bar{y}_{ij} = \frac{1}{n}\sum_{u=1}^{n} y_{iju} \tag{9.33}$$

Тогда

$$\bar{y}_i = \frac{1}{m}\sum_{j=1}^{m} \bar{y}_{ij}, \quad \bar{y}'_j = \frac{1}{k}\sum_{i=1}^{k} \bar{y}_{ij}, \tag{9.34}$$

$$\bar{y} = \frac{1}{km}\sum_{i=1}^{k}\sum_{j=1}^{m} \bar{y}_{ij} = \frac{1}{k}\sum_{i=1}^{k} \bar{y}_i \tag{9.35}$$

и дисперсии s_A^2 и s_B^2 рассчитываются по формулам (9.30) и (9.31).

В качестве оценки дисперсии воспроизводимости используем средневзвешенное значение дисперсий результатов в каждой ячейке

$$s_{\text{ош}}^2 = \frac{1}{mk}\sum_{i=1}^{k}\sum_{j=1}^{m} s_{ij}^2, \tag{9.36}$$

где

$$s_{ij}^2 = \frac{1}{n-1}\sum_{u=1}^{n}\left(y_{iju} - \bar{y}_{ij}\right)^2. \tag{9.37}$$

Число степеней свободы дисперсии $s_{\text{ош}}^2$ равно $f_{\text{ош}} = mk(n-1)$.

Введем также выборочную дисперсию, характеризующую влияние взаимодействия факторов

$$s_{AB}^2 \approx \frac{n}{(k-1)(m-1)}\left[\sum_{i=1}^{k}\sum_{j=1}^{m}\left(\bar{y}_{ij}-\bar{y}_i\right)^2 - k\sum_{j=1}^{m}\left(\bar{y}'_j-\bar{y}\right)^2\right], \quad (9.38)$$

с числом степеней свободы $f_{AB} = (k-1)(m-1)$.

Проверка значимости влияния факторов и их взаимодействия проводится по критерию Фишера, но неодинаково для моделей с фиксированными и случайными уровнями:

1. Для модели с фиксированными уровнями выборочные дисперсии s_A^2, s_B^2 и s_{AB}^2 сравниваются с оценкой дисперсии воспроизводимости $s_{\text{ош}}^2$. Если выполняются неравенства

$$\left(s_A^2 / s_{\text{ош}}^2\right) > F_{1-p}\left(f_A, f_{\text{ош}}\right), \quad \left(s_B^2 / s_{\text{ош}}^2\right) > F_{1-p}\left(f_B, f_{\text{ош}}\right),$$

$$\left(s_{AB}^2 / s_{\text{ош}}^2\right) > F_{1-p}\left(f_{AB}, f_{\text{ош}}\right), \quad (9.39)$$

то влияние факторов и их взаимодействия значимо.

2. Для модели со случайными уровнями проверка значимости взаимодействия факторов проводится так же, как и для для модели с фиксированными уровнями. Влияние факторов значимо, если выполняются следующие неравенства:

$$\left(s_A^2 / s_{AB}^2\right) > F_{1-p}\left(f_A, f_{AB}\right),$$

$$\left(s_B^2 / s_{AB}^2\right) > F_{1-p}\left(f_B, f_{AB}\right). \quad (9.40)$$

ЛЕКЦИЯ 10

Планирование эксперимента при дисперсионном анализе. Постановка задачи при планировании экстремальных экспериментов. Полный факторный эксперимент типа 2^2: матрица планирования, вычисление коэффициентов уравнения регрессии.

10.1. Планирование эксперимента при дисперсионном анализе

При двухфакторном дисперсионном анализе минимальное число опытов (в условиях линейной модели), обеспечивающее перебор всех возможных сочетаний уровней факторов, определяется произведением числа их уровней: $N = km$. Подобный эксперимент называется *полным факторным экспериментом* (ПФЭ). Если изучается влияние на процесс k факторов при одинаковом числе уровней n, то необходимое число опытов при ПФЭ равно

$$N = n^k . \tag{10.1}$$

Так, если $k = 2$ и $n = 3$ (табл. 3, лекция 9), то $N = 3^2 = 9$.

Эксперимент, в котором пропущены некоторые сочетания уровней, называется *дробным факторным экспериментом* (ДФЭ). Сокращение числа опытов неизбежно приводит к потере части информации, при этом обычно пренебрегают эффектами взаимодействия факторов.

Рассмотрим *трехфакторный дисперсионный анализ при одинаковом числе уровней n для каждого фактора*. Пусть $n = 2$. Тогда при ПФЭ потребуется провести $N = 2^3 = 8$ опытов (табл. 4).

Таблица 4

Полный факторный эксперимент 2^3

Уровни факторов	a_1		a_2	
	b_1	b_2	b_1	b_2
c_1	$^*y_{111}$	y_{121}	y_{211}	$^*y_{221}$
c_2	y_{112}	$^*y_{122}$	$^*y_{212}$	y_{222}

При отсутствии параллельных опытов результаты наблюдений можно представить в виде линейной модели

$$y_{ijq} = \mu + \alpha_i + \beta_j + \gamma_q + \varepsilon_{ijq}, \tag{10.2}$$

при этом линейные эффекты оказываются смешанными с эффектами взаимодействия: эффект A с BC взаимодействием, эффект B с AC взаимодействием, эффект C с AB взаимодействием. Однако число опытов

в условиях линейной модели можно существенно сократить при использовании ДФЭ, спланированного по схеме латинского квадрата.

Латинским квадратом $n \times n$ называют квадратную таблицу, составленную из n элементов (чисел или букв) таким образом, чтобы каждый элемент повторялся в каждой строке и каждом столбце только один раз. Из двух элементов образуется латинский квадрат 2 x 2:

$$\begin{matrix} A & B \\ B & A \end{matrix} \quad \text{или} \quad \begin{matrix} c_1 & c_2 \\ c_2 & c_1 \end{matrix}; \qquad (10.3)$$

из трех — латинский квадрат 3 x 3:

$$\begin{matrix} A & B & C \\ B & C & A \\ C & A & B \end{matrix} \quad \text{или} \quad \begin{matrix} c_1 & c_2 & c_3 \\ c_2 & c_3 & c_1 \\ c_3 & c_1 & c_2 \end{matrix}. \qquad (10.4)$$

Стандартными латинскими квадратами называются квадраты, у которых первые строка и столбец построены или в алфавитном порядке, или в порядке натурального ряда (квадраты (10.3) и (10.4)). Получены эти квадраты путем одношаговой циклической перестановки.

При ДФЭ по схеме латинского квадрата вводится в планирование третий фактор, при этом основой служит ПФЭ типа n^2. Так, при $n = 2$ на ПФЭ типа 2^2 (для факторов A и B) накладывается латинский квадрат 2 x 2 (табл. 5). План эксперимента, соответствующий табл. 5, называется *матрицей планирования* и представлен в табл. 6. Число опытов при этом сокращается до четырех вместо восьми при ПФЭ.

Таблица 5

2 x 2 латинский квадрат

A	B	
	b_1	b_2
a_1	c_1	c_2
a_2	c_2	c_1

Хотя латинский квадрат 2 x 2 является частью плана, всю табл. 5 также называют латинским квадратом. В нем каждый элемент повторяется только один раз в каждой строке и каждом столбце, что в равной степени сказывается при подсчете средних по строкам и столбцам. Приведенный в табл. 6 план представляет собой половину — *полуреплику* от ПФЭ типа 2^3 (вошедшие в полуреплику опыты отмечены в табл. 4 звездочками).

Таблица 6

План ДФЭ по схеме латинского квадрата 2 x 2
(k = 3, n = 2, N = 4)

Номер опыта	A	B	C	Итоги
1	a_1	b_1	c_1	y_{111}
2	a_1	b_2	c_2	y_{122}
3	a_2	b_1	c_2	y_{212}
4	a_2	b_2	c_1	y_{221}

Аналогично планируется ДФЭ по схеме латинского квадрата 3 x 3 (табл. 7). За основу взят ПФЭ типа 3^2, третий фактор (C) введен в рассмотрение по схеме латинского квадрата (10.4). ДФЭ 3^2 можно рассматривать как 1/3 реплику от ПФЭ типа 3^3.

Таблица 7

Латинский квадрат 3 x 3

A	B		
	b_1	b_2	b_3
a_1	c_1 y_1	c_2 y_2	c_3 y_3
a_2	c_2 y_4	c_3 y_5	c_1 y_6
a_3	c_3 y_7	c_1 y_8	c_2 y_9

В общем случае при планировании дробного факторного эксперимента по схеме латинского квадрата число опытов по сравнению с ПФЭ уменьшается в n раз (так, если $n = 4$, то при ПФЭ $N = 4^3 = 64$, а при ДФЭ по схеме латинского квадрата 4 x 4 — $N = 4^2 = 16$).

Дисперсионный анализ латинского квадрата, выполненного без параллельных опытов, проводится аналогично двухфакторному дисперсионному анализу. При этом для факторов A и B рассматривается их влияние на рассеяние средних по столбцам и по строкам относительно общего среднего соответственно, а для фактора C — на рассеяние средних по латинским буквам C_q. Так, например, для ДФЭ, представленного в табл. 7, средние по латинским буквам равны

$$C_1 = \frac{y_1 + y_6 + y_8}{3}, C_2 = \frac{y_2 + y_4 + y_9}{3}, C_3 = \frac{y_3 + y_5 + y_7}{3}. \qquad (10.5)$$

Значимость линейных эффектов проверяют по критерию Фишера. Адекватность принятой линейной модели можно проверить, выполнив для каждого сочетания уровней факторов (для каждой ячейки латинского квадрата) одинаковое число параллельных опытов. При этом наличие параллельных наблюдений используется только для оценки случайной ошибки опыта. Если эффекты взаимодействия незначимы, то остаточная дисперсия будет незначимо отличаться от дисперсии воспроизводимости, обусловленной ошибкой опыта.

10.2. Постановка задачи при планировании экстремальных экспериментов

Решение экстремальных задач физической химии и химической технологии (например, определение оптимальных условий проведения опыта и протекания процесса, оптимального состава материалов) возможно на основе математической модели объекта — *функции отклика*, связывающей выходной параметр, характеризующий результаты эксперимента, с переменными, определяющими условия проведения опыта (*факторами*):

$$y = \varphi\,(x_1, x_2, \ldots, x_k). \qquad (10.6)$$

На основе теоретического анализа физико-химических процессов при наличии достаточной информации об их механизмах можно составить детерминированную математическую модель объекта. Однако при проведении большинства исследований механизмы процессов, протекающих в изучаемых объектах, остаются неизвестными, поэтому для решения задач оптимизации необходимо использовать методы математической статистики.

При статистическом подходе математическая модель объекта или процесса представляется в виде полинома, т.е. отрезка ряда Тейлора, в который разлагается неизвестная функция (10.6):

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \sum_{\substack{u,j=1 \\ u \neq j}}^{k} \beta_{uj} x_u x_j + \sum_{j=1}^{k} \beta_{jj} x_j^2 +$$

$$+ \sum_{\substack{i,u,j=1 \\ i \neq u \neq j}}^{k} \beta_{iuj} x_i x_u x_j + \ldots \,, \qquad (10.7)$$

где

$$\beta_0 = \varphi(0),\ \beta_j = \frac{\partial \varphi(0)}{\partial x_j},\ \beta_{uj} = \frac{\partial \varphi(0)}{\partial x_u\, \partial x_j},$$

$$\beta_{jj} = \frac{\partial \varphi(0)}{2\partial x_j^2},\ \beta_{iuj} = \frac{\partial \varphi(0)}{\partial x_i\, \partial x_u\, \partial x_j}. \qquad (10.8)$$

Из-за воздействия случайных факторов на результаты опыта при обработке и анализе экспериментальных данных для полиномиальной модели (10.7) находят выборочные коэффициенты регрессии b_0, b_j, b_{uj}, b_{jj}, b_{uij}, которые являются оценками соответствующих теоретических коэффициентов. Уравнение регрессии записывается в виде

$$\hat{y} = b_0 + \sum_{j=1}^{k} b_j x_j + \sum_{\substack{u,j=1 \\ u \neq j}}^{k} b_{uj} x_u x_j + \sum_{j=1}^{k} b_{jj} x_j^2 + \sum_{\substack{i,u,j=1 \\ i \neq u \neq j}}^{k} b_{iuj} x_i x_u x_j + \ldots$$

$$+ \sum_{\substack{i,u,j=1 \\ i \neq u \neq j}}^{k} b_{iuj} x_i x_u x_j + \ldots, \qquad (10.9)$$

где b_0 — свободный член; b_j — линейные эффекты; b_{uj} — эффекты парного взаимодействия; b_{jj} — квадратичные эффекты; b_{uij} — эффекты тройного взаимодействия.

В зависимости от целей исследования и имеющейся информации можно ограничиться расчетом только части коэффициентов, пренебрегая влиянием остальных эффектов (например, в условиях линейной модели значимыми считаются только линейные эффекты, квадратичной модели — линейные и квадратичные эффекты, при этом в обоих случаях принимается, что эффекты взаимодействия факторов пренебрежимо малы).

Следует отметить, что на основании оценок теоретических коэффициентов нельзя определить аналитическое выражение функции отклика и, следовательно, получить информацию о механизме процесса. Полиномиальные модели используются только для решения задач оптимизации и управления процессами.

Под *планированием эксперимента* понимают оптимальное (наиболее эффективное) управление ходом эксперимента с целью получения максимально возможной информации на основе минимально допустимого количества опытных данных. Весь эксперимент обычно разбивается на несколько этапов. Информация, полученная после каждого этапа, используется для планирования исследований на следующем этапе. Планирование эксперимента позволяет варьировать все факто-

ры и получать одновременно количественные оценки всех эффектов, и при этом, в отличие от классического регрессионного анализа, избежать корреляции между коэффициентами уравнения регрессии.

10.3. Полный факторный эксперимент типа 2^2: матрица планирования, вычисление коэффициентов уравнения регрессии

При полном факторном эксперименте (ПФЭ) число опытов равно числу всех возможных комбинаций уровней факторов и при одинаковом числе уровней для каждого фактора определяется формулой

$$N = n^k, \tag{10.10}$$

где n — число уровней, k — число факторов ($j = 1, 2, \ldots, k$). ПФЭ 2^k называется такое проведение опытов, при котором каждый из k факторов рассматривается только на двух уровнях. При этом уровни факторов представляют собой границы варьирования данного параметра.

Допустим, что изучается влияние на выход продукта (y) двух параметров (факторов): температуры (z_1) в интервале 50–100 °C и давления (z_2) в диапазоне 1–2 атм. При реализации ПФЭ требуется выполнить $N = 2^2 = 4$ опыта. Произведем *кодирование факторов* (замену переменных):

$$x_j = \frac{z_j - z_j^{\text{o}}}{\Delta z_j}, \tag{10.11}$$

где

$$z_j^{\text{o}} = \frac{z_j^{\max} + z_j^{\min}}{2}, \; \Delta z_j = \frac{z_j^{\max} - z_j^{\min}}{2}, \tag{10.12}$$

$z_j^{\max}$ и $z_j^{\min}$ — верхняя и нижняя границы варьирования j-фактора. Точка $(z_1^{\text{o}}, z_2^{\text{o}})$ называется *центром плана*, или *основным уровнем*; величины Δz_1 и Δz_2 — интервалами варьирования по осям z_1 и z_2.

Как следует из уравнений (10.11) и (10.12), для переменных x_1 и x_2 нижний уровень равен –1, верхний — +1, координаты центра плана равны нулю. В табл. 8 представлен план ПФЭ 2^2, который в безразмерном масштабе может быть интерпретирован в виде четырех вершин квадрата (рис. 1).

Таблица 8

Полный факторный эксперимент 2^2

№ опыта	Факторы в натуральном масштабе		Факторы в безразмерном масштабе		Выход продукта, y
	z_1 (°C)	z_2 (атм)	x_1	x_2	
1	50	1	–1	–1	y_1
2	50	2	–1	+1	y_2
3	100	1	+1	–1	y_3
4	100	2	+1	+1	y_4

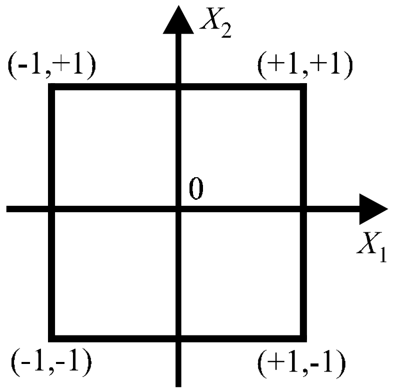

Рис. 1. Полный факторный эксперимент 2^2

Вычислим коэффициенты линейного уравнения регрессии

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 . \qquad (10.13)$$

Для нахождения b_0 в план ПФЭ надо ввести столбец фиктивной переменной $x_0 = 1$; соответствующая матрица планирования представлена в табл. 9. В математической статистике доказывается, что при планировании эксперимента по предложенной схеме и нахождении коэффициентов уравнения регрессии по методу наименьших квадратов любой коэффициент определяется скалярным произведением столбца y на соответствующий столбец факторов x_j в безразмерном масштабе (табл. 9), деленным на число опытов в матрице планирования:

$$b_j = \frac{\sum_{i=1}^{N} x_{ji} y_i}{N} . \qquad (10.14)$$

Таблица 9

Матрица планирования ПФЭ типа 2^2 с фиктивной переменной

№ опыта	x_0	x_1	x_2	y
1	+1	–1	–1	y_1
2	+1	–1	+1	y_2
3	+1	+1	–1	y_3
4	+1	+1	+1	y_4

Так, значение коэффицента b_1 определяется выражением

$$b_1 = \frac{1}{4}\sum_{i=1}^{4} x_{1i} y_i = \frac{[-y_1 + y_2 - y_3 + y_4]}{4}. \qquad (10.15)$$

Если ввести в рассмотрение эффект парного взаимодействия, то уравнение регрессии примет вид

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2. \qquad (10.16)$$

Для нахождения коэффициента b_{12} необходимо расширить матрицу планирования, представленную в табл. 9, добавив в нее столбец x_1x_2, характеризующий эффект взаимодействия (табл. 10).

Таблица 10

Расширенная матрица планирования ПФЭ типа 2^2

№ опыта	x_0	x_1	x_2	x_1x_2	y
1	+1	-1	-1	+1	y_1
2	+1	-1	+1	-1	y_2
3	+1	+1	-1	-1	y_3
4	+1	+1	+1	+1	y_4

Значения фактора взаимодействия в безразмерном масштабе определяются произведением соответствующих значений факторов x_1 и x_2:

$$(x_1 x_2)_i = x_{1i} \cdot x_{2i}. \qquad (10.17)$$

Коэффициент b_{12} определяется так же, как и линейные эффекты:

$$b_{12} = \frac{1}{N}\sum_{i=1}^{N} (x_1 x_2)_i y_i = \frac{1}{4}[y_1 - y_2 - y_3 + y_4]. \qquad (10.18)$$

ЛЕКЦИЯ 11

Матрица планирования ПФЭ 2^3. Проверка значимости коэффициентов и адекватности уравнения регрессии, полученных при обработке результатов ПФЭ 2^2 и 2^3. Дробный факторный эксперимент. Планы типа 2^{k-1}.

11.1. Матрица планирования полного факторного эксперимента типа 23

Рассмотрим планирование ПФЭ типа 2^3, при котором исследуется влияние на результат опыта уже трех факторов. При реализации такого ПФЭ требуется выполнить $N = 8$ опытов. Проведем кодирование факторов по уравнениям (10.11) – (10.12). План проведения опытов представлен в табл. 11, геометрически в безразмерном масштабе он может быть интерпретирован в виде восьми вершин куба (рис. 2).

Таблица 11

Полный факторный эксперимент 2^3

№ опыта	Факторы в безразмерном масштабе			Выход продукта, y
	x_1	x_2	x_3	
1	–1	–1	–1	y_1
2	+1	–1	–1	y_2
3	–1	+1	–1	y_3
4	+1	+1	–1	y_4
5	–1	–1	+1	y_5
6	+1	–1	+1	y_6
7	–1	+1	+1	y_7
8	+1	+1	+1	y_8

Уравнение регрессии с учетом эффектов взаимодействия факторов запишется в следующем виде:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 +$$
$$+ b_{23}x_2x_3 + b_{123}x_1x_2x_3, \qquad (11.1)$$

где коэффициенты b_{12}, b_{13} и b_{23} характеризуют эффекты парного взаимодействия, b_{123} — эффект тройного взаимодействия.

Для нахождения коэффициентов уравнения (11.1) необходимо составить расширенную матрицу планирования ПФЭ с фиктивной переменной, представленную в табл. 12.

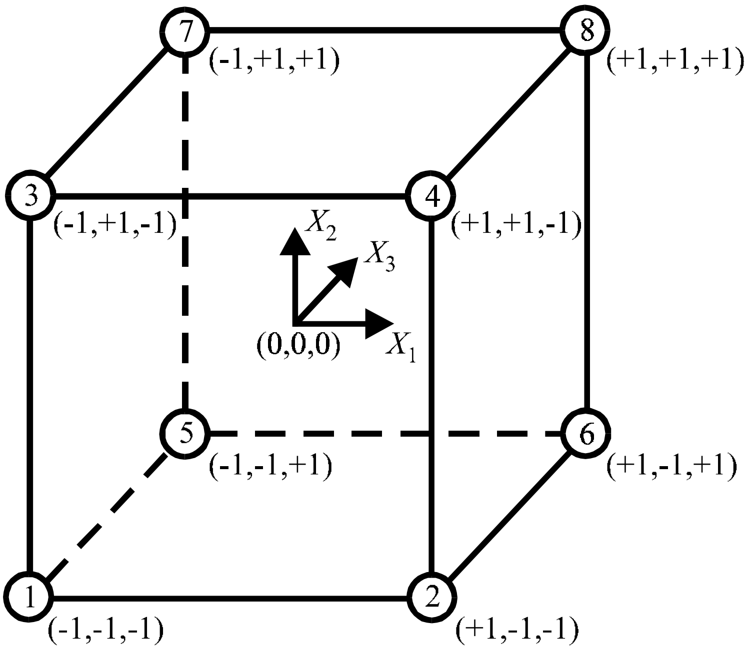

Рис. 2. Полный факторный эксперимент 2^3

Таблица 12

Расширенная матрица планирования ПФЭ типа 2^3

№	x_0	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$	y
1	+1	–1	–1	–1	+1	+1	+1	–1	y_1
2	+1	+1	–1	–1	–1	–1	+1	+1	y_2
3	+1	–1	+1	–1	–1	+1	–1	+1	y_3
4	+1	+1	+1	–1	+1	–1	–1	–1	y_4
5	+1	–1	–1	+1	+1	–1	–1	+1	y_5
6	+1	+1	–1	+1	–1	+1	–1	–1	y_6
7	+1	–1	+1	+1	–1	–1	+1	–1	y_7
8	+1	+1	+1	+1	+1	+1	+1	+1	y_8

Как и при ПФЭ 2^2, коэффициенты уравнения регрессии (11.1) определяются скалярным произведением столбца y на соответствующий столбец факторов или их взаимодействий в безразмерном масштабе, деленным на число опытов в матрице планирования (см. уравнения (10.14) и (10.18)).

Так, например, коэффициент b_{123} рассчитывается по следующему выражению:

$$b_{123} = \frac{1}{8}\left[-y_1 + y_2 + y_3 - y_4 + y_5 - y_6 - y_7 + y_8\right]. \qquad (11.2)$$

11.2. Проверка значимости коэффициентов и адекватности уравнения регрессии, полученных при обработке результатов ПФЭ 2^2 и 2^3

Для оценки значимости коэффициентов уравнения регрессии и проверки адекватности уравнения эксперименту достаточно провести серию параллельных опытов, выполненных при каком-то одном сочетании факторов.

Пусть в центре плана (в точках $(z_1^{\text{o}}, z_2^{\text{o}})$ и $(z_1^{\text{o}}, z_2^{\text{o}}, z_3^{\text{o}})$ для ПФЭ 2^2 и 2^3 соответственно) проведена серия из m опытов. Тогда выборочная дисперсия воспроизводимости, характеризующая влияние случайных факторов, равна

$$s^2_{\text{воспр}} = \frac{\sum\limits_{u=1}^{m} \left(y_u^{\text{o}} - \overline{y^{\text{o}}}\right)^2}{m-1}, \tag{11.3}$$

где y_u^{o} — результат u-го опыта ($u = 1, 2, \ldots, m$), $\overline{y^{\text{o}}}$ — среднее значение серии опытов. В математической статистике доказывается, что для спланированных экспериментов все коэффициенты уравнений регрессии определяются с одинаковой точностью, равной

$$s(b_j) = \frac{s_{\text{воспр}}}{\sqrt{N}}. \tag{11.4}$$

Значимость коэффициентов проверяется по критерию Стьюдента. В условиях нулевой гипотезы H_0: $\beta_j = 0$; отношение абсолютной величины коэффициента к его ошибке имеет распределение Стьюдента. Для каждого коэффициента определяется t-отношение:

$$t_j = \frac{|b_j|}{s(b_j)} = \frac{|b_j|}{s_{\text{воспр}}}\sqrt{N}, \tag{11.5}$$

которое сравнивается с табличным значением критерия Стьюдента $t_p(f)$ для выбранного уровня значимости p (обычно 0,05) и числа степеней свободы $f = m - 1$. Если для рассматриваемого коэффициента $t_j > t_p(f)$, то он значимо отличается от нуля. Выборочные коэффициенты, для которых $t_j \le t_p(f)$, незначимы, и их следует исключить из уравнения регрессии.

Допустим, при проверке значимости коэффициентов уравнения (11.1) оказалось, что все коэффициенты, характеризующие эффекты взаимодействия факторов, незначимы. После их исключения получаем линейное уравнение регрессии

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3, \qquad (11.6)$$

при этом значения b_0, b_1, b_2 и b_3 не требуется вычислять заново из-за того, что коэффициенты уравнения некоррелированы между собой. В отличие от классического регрессионного анализа, исключение незначимого коэффициента не сказывается на величинах остальных коэффициентов уравнения регрессии, а сами выборочные коэффициенты, полученные при реализации ПФЭ, являются *несмешанными* оценками теоретических коэффициентов.

Адекватность уравнения проверяется по критерию Фишера

$$F = \left(s_{\text{ад}}^2 \big/ s_{\text{воспр}}^2\right), \qquad (11.7)$$

Дисперсия адекватности (остаточная дисперсия) равна

$$s_{\text{ад}}^2 = s_{\text{ост}}^2 = \frac{1}{N-l}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \qquad (11.8)$$

где l — число значимых коэффициентов (для рассматриваемого случая $l = 4$). Уравнение адекватно описывает эксперимент, если

$$F \le F_{1-p}(f_1, f_2), \qquad (11.9)$$

где $F_{1\text{-}p}(f_1,f_2)$ — табличное значение критерия Фишера для $p = 0{,}05$ и чисел степеней свободы $f_1 = f_{\text{ад}} = N - l$ и $f_2 = f_{\text{воспр}} = m - 1$.

Рассмотрим также схему проведения регрессионного анализа для спланированного эксперимента в случае, когда каждый опыт в матрице планирования повторялся m раз. В качестве примера используем ПФЭ 2^3; при получении уравнения регрессии ограничимся линейным приближением (уравнение (11.6)). Матрица планирования такого эксперимента представлена в табл. 13.

Для каждого сочетания уровней факторов определяется среднее значение измеряемой величины и выборочная дисперсия:

$$\bar{y}_i = \frac{1}{m}\sum_{u=1}^{m} y_{iu}, \qquad (11.10)$$

$$s_i^2 = \frac{1}{m-1}\sum_{u=1}^{m}\left(y_{iu} - \bar{y}_i\right)^2. \qquad (11.11)$$

Таблица 13

Матрица планирования ПФЭ 2^3 в условиях линейной модели с одинаковым числом параллельных опытов при каждом сочетании уровней факторов

№	x_0	x_1	x_2	x_3	y	$\bar{y}_i$	s_i^2
1	+1	–1	–1	–1	$y_{11}, y_{12}, \ldots, y_{1m}$	$\bar{y}_1$	s_1^2
2	+1	+1	–1	–1	$y_{21}\, y_{22}, \ldots, y_{2m}$	$\bar{y}_2$	s_2^2
3	+1	–1	+1	–1	$y_{31}, y_{32}, \ldots, y_{3m}$	$\bar{y}_3$	s_3^2
4	+1	+1	+1	–1	$y_{41}, y_{42}, \ldots, y_{4m}$	$\bar{y}_4$	s_4^2
5	+1	–1	–1	+1	$y_{51}, y_{52}, \ldots, y_{5m}$	$\bar{y}_5$	s_5^2
6	+1	+1	–1	+1	$y_{61}, y_{62}, \ldots, y_{6m}$	$\bar{y}_6$	s_6^2
7	+1	–1	+1	+1	$y_{71}, y_{72}, \ldots, y_{7m}$	$\bar{y}_7$	s_7^2
8	+1	+1	+1	+1	$y_{81}, y_{82}, \ldots, y_{8m}$	$\bar{y}_8$	s_8^2

Однородность дисперсий проверяется по критерию Кохрена. Отношение максимальной дисперсии к сумме всех дисперсий

$$G = \frac{s_{\max}^2}{\sum_{i=1}^{N} s_i^2} \qquad (11.12)$$

сравнивается с табличным значением $G_{1-p}(f_1, f_2)$ для $p = 0{,}05$ и чисел степеней свободы $f_1 = m - 1$ и $f_2 = N$. Если $G \le G_{1-p}(f_1, f_2)$, то выборочные дисперсии однородны. Тогда наилучшей оценкой дисперсии воспроизводимости будет средневзвешенная дисперсия

$$s_{\text{воспр}}^2 = \frac{1}{N}\sum_{i=1}^{N} s_i^2 \qquad (11.13)$$

с числом степеней свободы $f_{\text{воспр}} = N(m - 1)$.

Коэффициенты уравнения регрессии определяются по формуле

$$b_j = \frac{1}{N}\sum_{i=1}^{N} x_{ji}\bar{y}_i\,. \qquad (11.14)$$

Поскольку дисперсия среднего в m раз меньше дисперсии единичного измерения, т. е.

$$s^2(\bar{y}) = s_{\text{воспр}}^2 / m\,, \qquad (11.15)$$

то выборочные среднеквадратичные отклонения коэффициентов рассчитываются следующим образом:

$$s(b_j) = \frac{s_{\text{воспр}}}{\sqrt{Nm}} = \frac{1}{N\sqrt{m}}\sqrt{\sum_{i=1}^{N} s_i^2}\,. \tag{11.16}$$

Значимость коэффициентов проверяется по критерию Стъюдента: если

$$t_j = \frac{|b_j|}{s(b_j)} > t_p(f), \tag{11.17}$$

где $t_p(f)$ — табличное значение критерия Стъюдента для $p = 0{,}05$ и числа степеней свободы $f = N(m-1)$, то коэффициент значимо отличается от нуля.

Адекватность уравнения регрессии эксперименту проверяется по критерию Фишера. Дисперсия адекватности равна

$$s_{\text{ад}}^2 = \frac{m\sum_{i=1}^{N}(\bar{y}_i - \hat{y}_i)^2}{N - l}, \tag{11.18}$$

где l — число значимых коэффициентов в уравнении регрессии.

Уравнение адекватно эксперименту, если

$$F = \frac{s_{\text{ад}}^2}{s_{\text{воспр}}^2} \le F_{1-p}(f_{\text{ад}}, f_{\text{воспр}}), \tag{11.19}$$

где $F_{1\text{-}p}(f_{\text{ад}}, f_{\text{воспр}})$ — табличное значение критерия Фишера для $p = 0{,}05$ и чисел степеней свободы $f_{\text{ад}} = N - l$ и $f_{\text{воспр}} = N(m-1)$. В противном случае для описания результатов эксперимента необходимо увеличить порядок аппроксимирующего полинома.

11.3. Дробный факторный эксперимент. Планы типа $2^{k\text{-}1}$

Число необходимых опытов в условиях линейной модели существенно сокращается при проведении дробных факторных экспериментов (дробных реплик от ПФЭ). В качестве реплики обычно используется полный факторный эксперимент для меньшего числа факторов. При этом вычисление коэффициентов уравнения и оценка их значимости проводится так же, как и в рассмотренных выше примерах ПФЭ 2^2 и 2^3. Число опытов в дробной реплике должно быть больше или равно числу неизвестных коэффициентов в уравнении регрессии.

Спланируем дробный факторный эксперимент для получения линейного уравнения регрессии небольшого участка поверхности отклика при трех независимых факторах:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 . \qquad (11.20)$$

Постановка ПФЭ 2^3 требует проведения 8 опытов. Для решения же поставленной задачи можно ограничиться 4 опытами, если в матрице планирования ПФЭ 2^2 (табл. 10, лекция 10) использовать столбец x_1x_2 в качестве плана для x_3. Матрица планирования такого сокращенного эксперимента — ДФЭ типа 2^{3-1}, или *полуреплики* от ПФЭ 2^3, — представлена в табл. 14.

Таблица 14

Матрица планирования ДФЭ типа 2^{3-1}

№ опыта	x_0	x_1	x_2	x_3	y
1	+1	–1	–1	+1	y_1
2	+1	–1	+1	–1	y_2
3	+1	+1	–1	–1	y_3
4	+1	+1	+1	+1	y_4

Проведение ДФЭ по предложенной схеме позволяет оценить свободный член и три коэффициента при линейных членах уравнения (11.20), однако при этом они будут являться несмешанными оценками теоретических коэффициентов только в том случае, если генеральные коэффициенты регрессии при парных взаимодействиях равны нулю. В противном случае найденные выборочные коэффициенты будут смешанными оценками теоретических:

$$b_1 \to \beta_1 + \beta_{23}, \quad b_2 \to \beta_2 + \beta_{13}, \quad b_3 \to \beta_3 + \beta_{12} . \qquad (11.21)$$

Генеральные коэффициенты не могут быть оценены по отдельности на основании только 4 опытов, поскольку при этом столбцы для линейных членов и парных произведений одинаковы (например, элементы вычисленного столбца для произведения x_2x_3 в точности совпадут с элементами столбца x_1). Чтобы определить, оценкой суммы каких именно генеральных коэффициентов явяются выборочные коэффициенты, удобно пользоваться *генерирующим соотношением*

$$x_3 = x_1 x_2 , \qquad (11.22)$$

в общем случае означающим, какой именно столбец ПФЭ 2^k был использован в качестве плана для введения $(k + 1)$-го фактора в ДФЭ. При умножении обоих частей (11.22) на x_3, получаем

$$x_3^2 = x_1 x_2 x_3. \qquad (11.23)$$

Единичный столбец

$$I = x_1 x_2 x_3 \qquad (11.24)$$

называется *определяющим контрастом* и позволяет определить, элементы каких столбцов в расширенной матрице планирования одинаковы. Умножая I по очереди на x_1, x_2 и x_3, получаем

$$x_1 = x_1^2 x_2 x_3 = x_2 x_3, \quad x_2 = x_1 x_2^2 x_3 = x_1 x_3,$$

$$x_3 = x_1 x_2 x_3^2 = x_1 x_2, \qquad (11.25)$$

в точности соответствующих системе смешанных оценок (11.21).

При постановке ДФЭ с числом факторов $k \geq 4$ в зависимости от генерирующего соотношения выборочные коэффициенты регрессии оказываются смешанными оценками того или иного сочетания генеральных коэффициентов. Поэтому важно заранее определиться с тем, какая информация является наиболее важной в данном исследовании, и в зависимости от поставленной задачи подобрать нужную дробную реплику.

Рассмотрим, например, планирование ДФЭ типа 2^{4-1}, представляющего собой полуреплику от ПФЭ 2^4. В качестве реплики используем ПФЭ 2^3 (табл. 12). Используем два генерирующих соотношения:

$$x_4 = x_1 x_2 x_3, \qquad (11.26)$$

$$x_4 = x_1 x_3. \qquad (11.27)$$

Для соотношения (11.26) определяющим контрастом будет

$$I = x_1 x_2 x_3 x_4. \qquad (11.28)$$

Тогда

$$x_1 = x_2 x_3 x_4, b_1 \to \beta_1 + \beta_{234}; \; x_2 = x_1 x_3 x_4, b_2 \to \beta_2 + \beta_{134};$$

$$x_3 = x_1 x_2 x_4, b_3 \to \beta_3 + \beta_{124}; \; x_4 = x_1 x_2 x_3, b_4 \to \beta_4 + \beta_{123};$$

$$x_1 x_2 = x_3 x_4, b_{12} \to \beta_{12} + \beta_{34}; \; x_1 x_3 = x_2 x_4, b_{13} \to \beta_{13} + \beta_{24};$$

$$x_1 x_4 = x_2 x_3, b_{14} \to \beta_{14} + \beta_{23}. \qquad (11.29)$$

В реальных задачах влияние тройных взаимодействий обычно равно нулю. Следовательно, генерирующее соотношение (11.26) следует использовать, если наибольший интерес представляют оценки для линейных эффектов.

Для соотношения (11.27) определяющим контрастом будет

$$I = x_1 x_3 x_4. \qquad (11.30)$$

Тогда

$$x_1 = x_3x_4,\, b_1 \to \beta_1 + \beta_{34};\;\; x_2 = x_1x_2x_3x_4,\, b_2 \to \beta_2 + \beta_{1234};$$

$$x_3 = x_1x_4,\, b_3 \to \beta_3 + \beta_{14};\;\; x_4 = x_1x_3,\, b_4 \to \beta_4 + \beta_{13};$$

$$x_1x_2 = x_2x_3x_4,\, b_{12} \to \beta_{12} + \beta_{234};\;\; x_2x_3 = x_1x_2x_4,\, b_{23} \to \beta_{23} + \beta_{124};$$

$$x_2x_4 = x_1x_2x_3,\, b_{24} \to \beta_{24} + \beta_{123}. \qquad (11.31)$$

Следовательно, дробную реплику с генерирующим соотношением $x_4 = x_1x_3$ следует использовать, если наибольший интерес представляют эффекты парных взаимодействий.

В общем случае число опытов в дробной реплике должно удовлетворять следующему соотношению:

$$k + 1 \le N < 2^k, \qquad (11.32)$$

где k — число факторов. Если число опытов равно числу определяемых коэффициентов в линейном уравнении регрессии ($N = k + 1$), дробная реплика представляет собой *линейный насыщенный план*, для которого все линейные эффекты смешаны с эффектами взаимодействия. Число степеней свободы остаточной дисперсии в таких планах равно нулю, поэтому для проверки адекватности линейного уравнения необходимо проведение дополнительных опытов.

Итак, рассмотренные двухуровневые планы ПФЭ 2^k и ДФЭ 2^{k-1} обладают следующими свойствами: вычисления просты; все коэффициенты регрессии определяются независимо друг от друга и с одинаковой и минимальной дисперсией; каждый коэффициент рассчитывается по результатам всех опытов.

ЛЕКЦИЯ 12

Оптимизация методом крутого восхождения по поверхности отклика. Описание функции отклика в области, близкой к экстремуму. Композиционные планы Бокса-Уилсона. Ортогональные планы второго порядка, расчет коэффицентов уравнения регрессии. Метод последовательного симплекс-планирования.

12.1. Оптимизация методом крутого восхождения по поверхности отклика

Задача оптимизации сводится к опытному определению такого сочетания уровней k факторов (координаты точки в (k+1)-мерном факторном пространстве), при котором достигается максимальное (минимальное) значение выходного параметра y (или нескольких параметров), т. е. функция отклика системы

$$y = \varphi(x_1, x_2, ..., x_k)$$

принимает экстремальное значение.

Рассмотрим случай, когда на систему оказывают влияние только два фактора (x_1 и x_2 в безразмерном масштабе). Построим контурные сечения $y = const$ поверхности отклика при $k = 2$ (рис. 3 *а*).

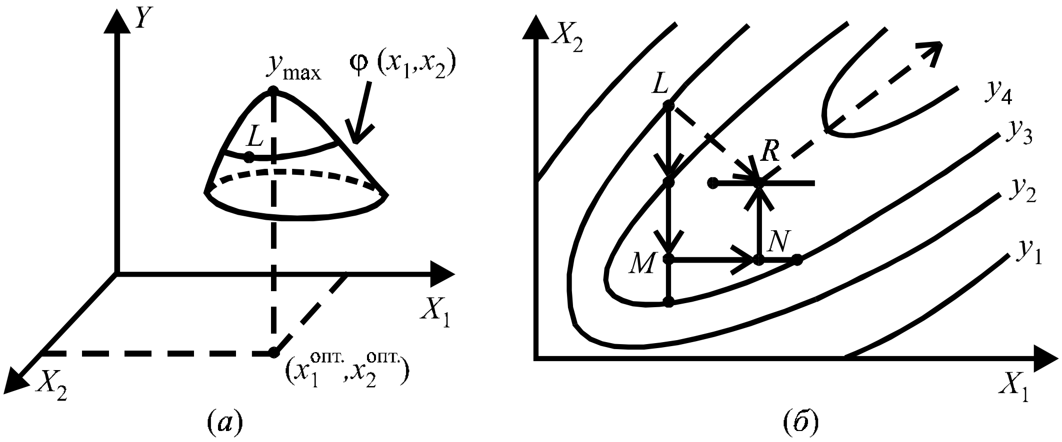

Рис. 3. Движение по поверхности отклика (*а*) к экстремуму в традиционном эксперименте и в методе крутого восхождения (*б*)

Поиск экстремальной точки поверхности отклика в традиционном эксперименте проводится следующим образом. В точке L с известным значением y фиксируется один из факторов, например x_1, и начинается движение из этой точки вдоль оси x_2. Движение по x_2 продолжается до тех пор, пока не прекращается прирост y (рис. 3 *б*). В точке M с наилучшим значением выходного параметра фиксируется фактор x_2 и начинается движение в направлении оси x_1. В точке N со следующим

наилучшим значением y снова фиксируется x_1 и начинается движение по x_2 и т. д. Очевидно, что путь к экстремуму по ломаной кривой *LMNR* (рис. 3 *б*) не является оптимальным.

Кратчайшим, наиболее крутым путем достижения экстремума будет движение из точки L по градиенту перпендикулярно изолиниям $y = const$ (на рис. 3 *б* этот путь показан пунктирной линией). Для рассматриваемого случая градиент функции отклика равен

$$\operatorname{grad} \varphi = \left(\frac{\partial \varphi}{\partial x_1} \right) \vec{i} + \left(\frac{\partial \varphi}{\partial x_2} \right) \vec{j}, \tag{12.1}$$

где $\vec{i}$ и $\vec{j}$ — орты координатных осей. Предполагается, что функция φ непрерывна, дифференцируема и не имеет особых точек.

Для реализации метода крутого восхождения Бокс и Уилсон предложили шаговый метод движения по поверхности отклика. В окрестности точки L ставится эксперимент для локального описания поверхности отклика линейным уравнением регрессии:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 . \tag{12.2}$$

Движение из точки L начинается в направлении градиента линейного приближения

$$\frac{\partial \hat{y}}{\partial x_1} = b_1; \quad \frac{\partial \hat{y}}{\partial x_2} = b_2. \tag{12.3}$$

Для случая, представленного на рис. 3 *б*, выборочные коэффициенты при линейных членах в окрестности точки L имеют разные знаки: $b_1 > 0$, $b_2 < 0$, поэтому при движении к максимуму функции отклика значение x_1 увеличивается, а x_2 уменьшается. Движение по градиенту линейного приближения продолжается до тех пор, пока не прекращается прирост y. В точке с наибольшим значением y (центр плана) ставится новая серия опытов и определяется новое направление движения по поверхности отклика. Такой шаговый процесс продолжается до достижения области, близкой к экстремуму.

При постановке опытов величина шага должна быть пропорциональна произведению коэффициента на интервал варьирования: $b_j \Delta z_j$. Например, при движении из точки L следующий эксперимент ставится в точке со значениями x_1 и x_2, отличающимися от начальных на величины $2b_1\Delta z_1$ и $2b_2\Delta z_2$ соответственно. В общем случае направление градиента будет зависеть от выбранного интервала варьирования независимых факторов. При изменении в n раз интервала варьирования некоторого j-фактора величина шага для него меняется в n^2 раз, так

как при этом в n раз изменяется и коэффициент регрессии b_j. Инвариантными к изменению интервала остаются только знаки составляющих градиента. При увеличении числа рассматриваемых факторов более двух оптимизация методом крутого восхождения по поверхности отклика проводится аналогичным способом.

12.2. Описание функции отклика в области, близкой к экстремуму. Композиционные планы Бокса-Уилсона

В области, близкой к экстремуму, (или «почти стационарной области») функция отклика существенно нелинейна, поэтому для ее адекватного описания необходимо использовать нелинейные полиномы. В настоящее время для этой цели наиболее широко применяют полиномы второго порядка, для получения которых имеются хорошо разработанные планы эксперимента.

Для описания полиномом второго порядка эксперимента, реализованного для нахождения оптимальных условий процесса, число опытов N в плане должно быть не меньше числа определяемых коэффициентов в уравнении регрессии второго порядка для k факторов

$$\hat{y} = b_0 + \sum_{j=1}^{k} b_j x_j + \sum_{\substack{u,j=1 \\ u \neq j}}^{k} b_{uj} x_u x_j + \sum_{j=1}^{k} b_{jj} x_j^2 . \qquad (12.4)$$

Выборочные коэффициенты (12.4) являются оценками соответствующих коэффициентов уравнения теоретической регрессии:

$$m_y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \sum_{\substack{u,j=1 \\ u \neq j}}^{k} \beta_{uj} x_u x_j + \sum_{j=1}^{k} \beta_{jj} x_j^2 . \qquad (12.5)$$

В зависимости от числа рассматриваемых факторов число коэффициентов l уравнения регрессии (12.4) определяется по формуле

$$l = (k+1) + k + C_k^2 = 2k + 1 + \frac{k!}{2!(k-2)!} = \frac{(k+1)(k+2)}{2}, \qquad (12.6)$$

где C_k^2 — количество сочетаний из k факторов по два, равное числу эффектов парного взаимодействия.

В области, близкой к экстремуму, становятся значимыми эффекты парного взаимодействия и квадратичные эффекты. Поэтому то, что адекватное описание результатов эксперимента требует использования полиномов второго порядка, может служить признаком нахождения в почти стационарной области. Близость к этой области можно

также установить, поставив дополнительно к ПФЭ 2^k или ДФЭ 2^{k-1} серию опытов в центре плана. Среднее значение результатов этих опытов является оценкой для свободного члена уравнения (12.5):

$$\overline{y^{\text{o}}} \to \beta_0 . \tag{12.7}$$

Выборочный коэффицент b_0, вычисляемый по формуле

$$b_0 = \frac{1}{N}\sum_{i=1}^{N} x_{0i} y_i , \tag{12.8}$$

оценивает сумму свободного и квадратичных членов:

$$b_0 \to \beta_0 + \sum_{j=1}^{k} \beta_{jj} . \tag{12.9}$$

Поэтому, чем больше разность

$$(b_0 - \overline{y^{\text{o}}}) \to \sum_{j=1}^{k} \beta_{jj} , \tag{12.10}$$

тем значимее квадратичные эффекты.

Для описания поверхности отклика полиномами второго порядка независимые факторы в планах должны принимать не менее трех разных значений. Эксперимент, в котором каждый из k факторов рассматривается на трех уровнях и реализуются все возможные сочетания уровней факторов, является ПФЭ типа 3^k. В качестве примера в табл. 15 представлена матрица планирования ПФЭ 3^2.

Проведение ПФЭ 3^k требует большого числа опытов, намного превышающего число определяемых коэффициентов l в уравнении (12.4) уже при $k > 2$:

k	2	3	4
3^k	9	27	81
l	6	10	15

Таблица 15

Матрица планирования ПФЭ 3^2

№ опыта	x_1	x_2	y
1	–1	–1	y_1
2	0	–1	y_2
3	+1	–1	y_3
4	–1	0	y_4
5	0	0	y_5
6	+1	0	y_6
7	–1	+1	y_7
8	0	+1	y_8
9	+1	+1	y_9

Сократить общее число опытов при условии получения несмешанных оценок для линейных эффектов и эффектов взаимодействия можно с помощью *композиционных планов Бокса-Уилсона*. Ядро таких планов при $k < 5$ составляет ПФЭ 2^k, и полуреплика от него при $k \geq 5$.

Если линейное уравнение регрессии оказалось неадекватным эксперименту, необходимо:

1) добавить $2k$ звездных точек, расположенных на координатных осях факторного пространства. Координаты звездных точек в общем случае равны

$$(\pm\alpha, 0, ..., 0), (0, \pm\alpha, ..., 0), ..., (0, ..., 0, \pm\alpha),$$

где α — расстояние от центра плана до звездной точки, или звездное плечо;

2) увеличить число экспериментов в центре плана n_0.

Общее число опытов в матрице композиционного плана при $k \leq 4$ составляет

$$N = 2^k + 2k + n_0. \tag{12.11}$$

Рассмотрим построение композиционных планов на примере $k = 2$ (рис. 4). Точки 1, 2, 3, 4 образуют ПФЭ 2^2, точки 5, 6, 7, 8 являются звездными точками с координатами $(\pm\alpha, 0)$ и $(0, \pm\alpha)$, координаты n_0 опытов в центре плана нулевые — (0, 0).

Композиционный план второго порядка для двух факторов представлен в табл. 16, при этом в центре плана выполнена серия из трех опытов (№ 9–11).

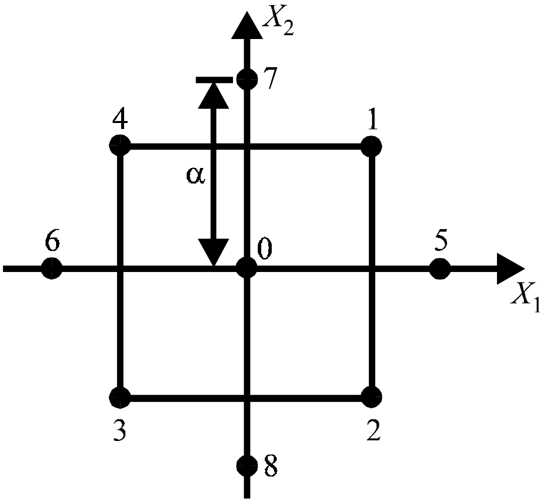

Рис. 4. Композиционный план второго порядка для двух факторов

Таблица 16

Композиционный план второго порядка для двух факторов

№ опыта	x_0	x_1	x_2	x_1x_2	x_1^2	x_2^2	y
1	+1	+1	+1	+1	+1	+1	y_1
2	+1	+1	–1	–1	+1	+1	y_2
3	+1	–1	–1	+1	+1	+1	y_3
4	+1	–1	+1	–1	+1	+1	y_4
5	+1	+α	0	0	α^2	0	y_5
6	+1	–α	0	0	α^2	0	y_6
7	+1	0	+α	0	0	α^2	y_7
8	+1	0	–α	0	0	α^2	y_8
9	+1	0	0	0	0	0	y_9
10	+1	0	0	0	0	0	y_{10}
11	+1	0	0	0	0	0	y_{11}

12.3. Ортогональные планы второго порядка, расчет коэффициентов уравнения регрессии

Выбор звездного плеча в композиционных планах Бокса–Уилсона может быть произвольным, однако расчеты коэффициентов уравнения регрессии при $k < 5$ существенно упрощаются, если величина плеча определяется исходя из следующего уравнения:

$$\alpha^4 + 2k\alpha^2 - 2^{k-1}(k + 0.5n_0) = 0 . \qquad (12.12)$$

Значения α^2, определенные по (12.12), приведены в табл. 17.

Таблица 17

Значения α^2 для k факторов и n_0 опытов в центре плана

n_0	k			n_0	k		
	2	3	4		2	3	4
1	1.00	1.476	2.00	6	1.742	2.325	2.950
2	1.160	1.650	2.164	7	1.873	2.481	3.140
3	1.317	1.831	2.390	8	2.00	2.633	3.310
4	1.475	2.00	2.580	9	2.113	2.782	3.490
5	1.606	2.164	2.770	10	2.243	2.928	3.66

Выбрав α, проведем следующее линейное преобразование квадратичных столбцов:

$$x'_j = x_j^2 - \overline{x_j^2} = x_j^2 - \frac{1}{N}\sum_{i=1}^{N} x_{ji}^2 . \qquad (12.13)$$

Композиционные планы, полученные таким образом, называются *ортогональными планами второго порядка*.

Ортогональный план второго порядка при $k = 2$ и $n_0 = 1$ представлен в табл. 18. За его основу взят композиционный план для двух факторов (табл. 16) с общим числом опытов $N = 9$. Величину звездного плеча определим по табл. 17: $\alpha^2 = 1$, $\alpha = 1$. Средние значения элементов квадратичных столбцов в табл. 16 равны

$$\overline{x_1^2} = \frac{1}{9}\sum_{i=1}^{9} x_{1i}^2 = \overline{x_2^2} = \frac{1}{9}\sum_{i=1}^{9} x_{2i}^2 = \frac{1}{9}(4 + 2\acute{a}^2) = \frac{2}{3}. \qquad (12.14)$$

В математической статистике доказывается, что для ортогональных планов второго порядка все коэффициенты уравнения регрессии определяются независимо друг от друга по формуле

$$b_j = \sum_{i=1}^{N} x_{ji} y_i \Bigg/ \sum_{i=1}^{N} x_{ji}^2 , \qquad (12.15)$$

а дисперсии коэффициентов равны

$$s^2(b_j) = s_{\text{воспр}}^2 \Bigg/ \sum_{i=1}^{N} x_{ji}^2 . \qquad (12.16)$$

Таблица 18

Ортогональный план второго порядка для двух факторов

№ опыта	x_0	x_1	x_2	x_1x_2	x_1'	x_2'	y
1	+1	+1	+1	+1	+1/3	+1/3	y_1
2	+1	+1	–1	–1	+1/3	+1/3	y_2
3	+1	–1	–1	+1	+1/3	+1/3	y_3
4	+1	–1	+1	–1	+1/3	+1/3	y_4
5	+1	+α	0	0	+1/3	–2/3	y_5
6	+1	–α	0	0	+1/3	–2/3	y_6
7	+1	0	+α	0	–2/3	+1/3	y_7
8	+1	0	–α	0	–2/3	+1/3	y_8
9	+1	0	0	0	–2/3	–2/3	y_9

Для определения дисперсии воспроизводимости необходимо выполнить серию опытов в центре плана. В результате расчетов по матрице с преобразованными столбцами для квадратичных эффектов (табл. 18) получаем следующее уравнение:

$$\hat{y} = b_0' + b_1x_1 + b_2x_2 + b_{12}x_1x_2 + b_{11}(x_1^2 - \overline{x_1^2}) + b_{22}(x_2^2 - \overline{x_2^2}). \quad (12.17)$$

Чтобы перейти к обычной записи уравнения регрессии в виде

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2 + b_{11}x_1^2 + b_{22}x_2^2, \quad (12.18)$$

определим b_0 по формуле

$$b_0 = b_0' - b_{11}\overline{x_1^2} - b_{22}\overline{x_2^2} \quad (12.19)$$

с дисперсией, равной

$$s^2(b_0) = s^2(b_0') + \left(\overline{x_1^2}\right)^2 s^2(b_{11}) + \left(\overline{x_2^2}\right)^2 s^2(b_{22}). \quad (12.20)$$

Значимость коэффициентов проверяется по критерию Стьюдента. Если

$$t_j = \frac{|b_j|}{s(b_j)} > t_p(f), \quad (12.21)$$

где $t_p(f)$ — табличное значение критерия Стьюдента для $p = 0{,}05$ и числа степеней свободы дисперсии воспроизводимости, то коэффициент значимо отличается от нуля.

Коэффициенты уравнения регрессии, получаемые при помощи ортогональных планов второго порядка, определяются с разной точностью. В случае, когда $k \le 4$, согласно (12.16) имеем

$$s^2(b_0') = \frac{s^2_{\text{воспр}}}{N};$$

$$s^2(b_j) = \frac{s^2_{\text{воспр}}}{2^k + 2\alpha^2},\ j = 1, 2, ..., k\ ;$$

$$s^2(b_{uj}) = \frac{s^2_{\text{воспр}}}{2^k}, u, j = 1, 2, ..., k, u \ne j; \tag{12.22}$$

$$s^2(b_{jj}) = \frac{s^2_{\text{воспр}}}{2^k(1-\overline{x_j^2})^2 + 2(\alpha^2 - \overline{x_j^2})^2 + (n_0 + 2k - 2)(\overline{x_j^2})^2},\ j = 1, 2, ..., k.$$

После исключения незначимых коэффициентов проводится проверка адекватности уравнения по критерию Фишера. Уравнение адекватно эксперименту, если

$$F = \frac{s^2_{\text{ад}}}{s^2_{\text{воспр}}} \le F_{1-p}(f_{\text{ад}}, f_{\text{воспр}}), \tag{12.23}$$

где $F_{1\text{-}p}(f_{\text{ад}},f_{\text{воспр}})$ — критерий Фишера для $p = 0{,}05$; $f_{\text{ад}} = N - l$ — число степеней свободы дисперсии адекватности (l — число значимых коэффициентов в уравнении регрессии); $f_{\text{воспр}}$ — число степеней свободы дисперсии воспроизводимости.

12.4. Метод последовательного симплекс-планирования

В рассмотренных выше планах ПФЭ 2^2 и 2^3 экспериментальные точки располагались в вершинах квадрата и куба соответственно. В качестве экспериментального плана можно также использовать *регулярный симплекс*. Симплексом в k-мерном пространстве называют выпуклый многогранник, имеющий ровно $(k + 1)$ вершину, каждая из которых определяется пересечением k гиперплоскостей данного пространства. Симплекс называется регулярным, если расстояния между всеми его вершинами равны. Примерами регулярных симплексов являются правильный треугольник в двумерном пространстве и тетраэдр в трехмерном.

На практике планирование эксперимента с использованием регулярных симплексов применяется для решения задач оптимизации при

движении к почти стационарной области. Для получения регулярного симплекса проводится линейное преобразование уровней факторов

$$x_j = \frac{z_j - z_j^{o}}{\Delta z_j}, \qquad (12.24)$$

где z_j^{o} — j-я координата центра плана; Δz_j — интервал варьирования по j-фактору.

Оптимизация методом последовательного симплекс-планирования проводится следующим образом: исходная серия опытов планируется так, чтобы экспериментальные точки образовывали регулярный симплекс в факторном пространстве. После проведения опытов определяется вершина симплекса, соответствующая наихудшим результатам. Далее строится новый симплекс, для чего наихудшая точка исходного симплекса заменяется новой, расположенной симметрично относительно центра грани симплекса, находящейся против наихудшей точки. Новая точка вместе с оставшимися точками образует новый симплекс, центр тяжести которого смещен в сторону повышения качества процесса. После реализации опыта в дополнительной точке опять проводится выявление наихудшей вершины симплекса и т. д. При достижении области оптимума, симплекс начинает вращение вокруг вершины с максимальным значением отклика.

На рис. 5 показаны схемы достижения экстремума поверхности отклика методами крутого восхождения и симплекс-планирования на примере зависимости целевой функции y от двух факторов. При оптимизации методом крутого восхождения (рис. 5 *а*) в окрестности точки M поставлен ПФЭ 2^2, движение по градиенту линейного приближения осуществлялось в опытах 5–9. Далее был поставлен новый ПФЭ 2^2 (точки 10–13) с центром в точке 7, в которой было получено наилучшее значение y. Движение по новому градиенту (точки 14–15) приводит к экстремуму.

При оптимизации методом симплекс-планирования (рис. 5 *б*) в исходном симплексе (точки 1–3) худшей точкой оказалась точка 2. Ее зеркальным отражением относительно c_1 — центра грани 1–3 — является точка 4. В новом симплексе 1, 3, 4 худшей оказалась точка 1, в результате ее зеркального отражения получен симплекс 3, 4, 5 и т. д. Область оптимума достигается при реализации симплекса 9, 10, 11. Хотя оба рассмотренных метода требуют проведения примерно одинакового числа опытов, симплекс-планирование имеет ряд важных

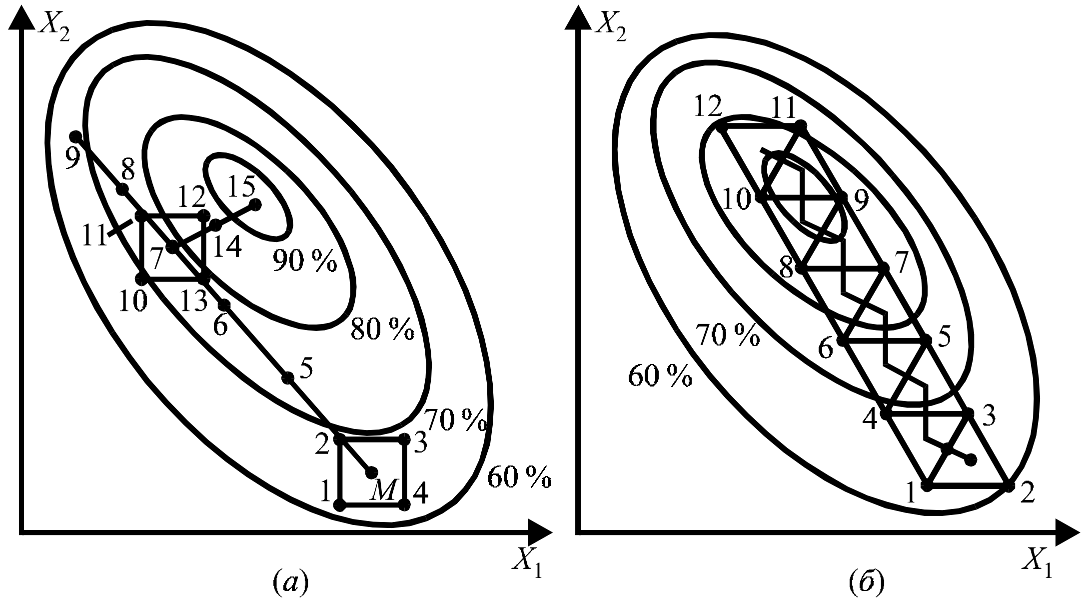

Рис. 5. Достижение экстремума поверхности отклика методами крутого восхождения (*а*) и симплекс-планирования (*б*)

преимуществ: при использовании этого метода параметр оптимизации y может измеряться приближенно, достаточно иметь возможность проранжировать его величину; можно одновременно учитывать несколько параметров оптимизации; метод не предъявляет жестких требований к локальной аппроксимации поверхности отклика уравнением регрессии.

На практике рекомендуется ориентировать исходный симплекс в факторном пространстве следующим образом: центр симплекса совпадает с началом координат, одна из вершин лежит на координатной оси, а остальные располагаются симметрично относительно координатных осей, плоскостей и гиперплоскостей (в многомерном случае). Тогда координаты вершин симплекса при $k = 5$ задаются матрицей

$$X = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ -x_1 & x_2 & x_3 & x_4 & x_5 \\ 0 & -2x_2 & x_3 & x_4 & x_5 \\ 0 & 0 & -3x_3 & x_4 & x_5 \\ 0 & 0 & 0 & -4x_4 & x_5 \\ 0 & 0 & 0 & 0 & -5x_5 \end{bmatrix}. \qquad (12.25)$$

При $k < 5$ координаты вершин симплекса определяются частью матрицы (12.25), при этом число столбцов равно числу факторов, а число строк равно $k + 1$; при $k > 5$ для каждого добавленного фактора в матрицу (12.25) добавляются соответствующие столбец и строка. В об-

щем случае число опытов в симплексной матрице для k независимых факторов $N = (k + 1)$ равно числу коэффициентов линейного уравнения регрессии, т. е. симплексные планы являются насыщенными.

Если длину стороны симплекса принять равной 1, то

$$x_j = \sqrt{\frac{1}{2j(j+1)}}. \tag{12.26}$$

Для практического использования матрицы (12.25) ее числовые элементы заранее подсчитаны по формуле (12.26)

$$X = \begin{bmatrix} 0.5 & 0.289 & 0.204 & 0.158 & 0.129 \\ -0.5 & 0.289 & 0.204 & 0.158 & 0.129 \\ 0 & -0.578 & 0.204 & 0.158 & 0.129 \\ 0 & 0 & -0.612 & 0.158 & 0.129 \\ 0 & 0 & 0 & -0.632 & 0.129 \\ 0 & 0 & 0 & 0 & -0.645 \end{bmatrix}. \tag{12.27}$$

План эксперимента в безразмерном масштабе для k факторов состоит из k столбцов и $k + 1$ строки матрицы (12.27). Коэффициенты уравнения линейной регрессии

$$\hat{y} = b_0 + \sum_{j=1}^{k} b_j x_j \tag{12.28}$$

вычисляются следующим образом:

$$b_0 = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad b_j = 2\sum_{i=1}^{N} x_{ji} y_i. \tag{12.29}$$

Если в одной из вершин симплекса поставить серию параллельных опытов и рассчитать дисперсию воспроизводимости, то выборочные дисперсии коэффициентов определяются по формуле

$$s^2(b_j) = \frac{s^2_{\text{воспр}}}{\sum_{i=1}^{N} x_{ji}^2} = 2s^2_{\text{воспр}}. \tag{12.30}$$

Следует отметить, что коэффициенты уравнения регрессии, полученные по симплексному плану, определяются с меньшей точностью по сравнению с коэффицентами, полученными при реализации ПФЭ 2^k и ДФЭ $2^{k\text{-}1}$, для которых

$$s^2(b_j) = s^2_{\text{воспр}} / N. \quad (12.31)$$

После реализации исходного симплекса требуется провести отражение наихудшей точки относительно центра противоположной грани. Координаты отраженной точки равны:

$$x_j^{(k+2)} = 2x_j^{(c)} - x_j^{(l)}, \quad j = 1, 2, ..., k, \quad (12.32)$$

где $x_j^{(l)}$ — j-я координата наихудшей точки; $x_j^{(k+2)}$ — j-я координата новой точки, получаемой в результате отражения; $x_j^{(c)}$ — j-я координата центра противоположной грани, определяемая по формуле

$$x_j^{(c)} = \frac{1}{k}\sum_{i=1}^{k+1} x_j^{(i)}, \quad i \neq l, \quad (12.33)$$

где $x_j^{(i)}$ — j-я координата i-й вершины симплекса ($i = 1, 2, \ldots, k+1$).

Координаты центра оптимального симплекса (точка S) в почти стационарной области находятся следующим образом:

$$x_j^{(S)} = \frac{1}{(k+1)}\sum_{i=1}^{k+1} x_j^{(i)}. \quad (12.34)$$

Учебное издание

Блохин Андрей Викторович

ТЕОРИЯ ЭКСПЕРИМЕНТА

КУРС ЛЕКЦИЙ

В двух частях

Часть 2

В авторской редакции

Художник обложки А. *А. Федорченки*
Технический редактор Г. *М. Романчук*
Корректор *Р. П. Кадырко*

Ответственный за выпуск *А. В. Блохин*

Подписано в печать 20.11.2002. Формат 60x84/16. Бумага офсетная.
Гарнитура Тайме. Печать офсетная Уел печ. л. 3,95. Уч.-изд. л. 3,62.
Тираж 100 экз. Зак.58

Белорусский государственный университет.
Лицензия ЛВ №315 от 14.07.98.
220050, Минск, проспект Франциска Скорины, 4.

Отпечатано с оригинала-макета заказчика.
Республиканское унитарное предприятие
«Издательский центр Белорусского государственного университета».
Лицензия ЛП№ 461 от 14.08.2001.
220030, Минск, ул. Красноармейская, 6.